Psychological Monographs: General and Applied

Combining the *Applied Psychology Monographs* and the *Archives of Psychology* with the *Psychological Monographs*

VOL. 80

1966

GREGORY A. KIMBLE, Editor
Duke University
Durham, North Carolina

Published by
THE AMERICAN PSYCHOLOGICAL ASSOCIATION, INC.
1200 SEVENTEENTH STREET, N. W., WASHINGTON, D. C. 20036

CONTENTS OF VOLUME 80

Whole No.

Vol. 80, No. 3 Whole No. 611, 1966

Psychological Monographs: General and Applied

UNDERCONTROLLED AND OVERCONTROLLED PERSONALITY TYPES IN EXTREME ANTISOCIAL AGGRESSION[1]

EDWIN I. MEGARGEE[2]

University of California, Berkeley

Physical aggression is typically attributed to inadequate control. While this is the pattern in 1 type of physically aggressive person, it is proposed that in another type, the Chronically Overcontrolled, rigid inhibitions against overt aggressive behavior will be found. Aggression by such people is apt to be of murderous intensity as aggressive impulse must build up to higher levels to overcome such inhibitions and since alternative means of expressing aggression have not been learned. This suggests that in comparison with other criminal groups, a murderously assaultive group will be assessed as less hostile, less aggressive, and more controlled. An empirical study of 4 groups of assaultive and nonviolent delinquents supports this prediction. Implications of this finding for practice and theory are discussed.

Aggression and violence are more than ever becoming major problems in the United States. The names of Los Angeles, Rochester, and St. Augustine have joined Bunker Hill, Gettysburg, and the Little Big Horn as American battlegrounds. There is also concern regarding individual violence. In a single week a national magazine reported the cases of two 22-year-old boys, one (a "gentle, easy-going, good natured" young man) who 5 days after graduation killed three unarmed victims during a bank robbery and the other (a "mild and loving" person) who shot his twin brother (*Newsweek*, 1965).

When we try to apply information gleaned from empirical studies of aggression to events such as these we find a great gap between the aggression described in our journals and that described in our newspapers. Most empirical data have been collected either in the laboratory under controlled conditions or in the schoolyard using the method of naturalistic observation. In either case, the amount of extreme aggression that can take place is seriously curtailed, either by the experimenter's ethics or by the intervention of school personnel. For this reason most of our data concern relatively mild forms of aggression and the psychologist must extrapolate to account for more extreme aggression such as assault or homicide.

The general formulation that has emerged from empirical studies of relatively mild aggression is that the overtly aggressive person has fewer controls and more need or instigation for aggression than does the overtly nonaggressive person.

The practical implications of this are clear: the way to discourage a person from acting aggressively is to build up his controls. Our prisons and reformatories typically base their programs upon this principle by instituting rewards for control and punishments for aggression. When an individual has demonstrated his controls by behaving in a nonaggressive fashion for a sufficiently long period, he is considered to be rehabilitated and is considered for release.

However, there is reason to believe the dynamics underlying an extremely assaultive offense such as homicide may be quite

[1] Based on a doctoral dissertation submitted to the Department of Psychology at the University of California, Berkeley. The author wishes to express his appreciation to the members of his doctoral committee, Hubert Coffey and Irving Piliavin, and, most especially, to the chairman, Gerald A. Mendelsohn, for assistance in the design, execution, and interpretation of the study. He also wishes to thank Lorenzo S. Buckley, Chief Probation Officer of Alameda County, California, and the staffs of the Guidance Clinic and of Juvenile Hall who collected and transcribed the data. Finally, he wishes to thank the University of California Computer Center for donating time on the IBM 7090.

[2] Now at the University of Texas.

different from the dynamics found in milder aggressive behavior.[3] In case after case the extremely assaultive offender proves to be a rather passive person with no previous history of aggression. In Phoenix an 11-year-old boy who stabbed his brother 34 times with a steak knife was described by all who knew him as being extremely polite and soft spoken with no history of assaultive behavior. In New York an 18-year-old youth who confessed he had assaulted and strangled a 7-year-old girl in a Queens church and later tried to burn her body in the furnace was described in the press as an unemotional person who planned to be a minister. A 21-year-old man from Colorado who was accused of the rape and murder of two little girls had never been a discipline problem and, in fact, his stepfather reported, "When he was in school the other kids would run all over him and he'd never fight back. There is just no violence in him." In these cases the homicide was not just one more aggressive offense in a person who had always displayed inadequate controls, but rather a completely uncharacteristic act in a person who had always displayed extraordinarily high levels of control.

There are empirical as well as anecdotal data which indicate that extreme and moderate aggressive behavior might be characterized by different dynamics. For instance, in a study of the MMPI in which hostility scale scores of assaultive and nonassaultive criminals were compared, Megargee and Mendelsohn (1962) found a pattern of reversals with the assaultive subjects being tested as having more control and less hostility than the nonassaultive criminals or normals. This led them to suggest

> that the extremely assaultive person is often a fairly mild-mannered, long-suffering individual who buries his resentment under rigid but brittle controls. Under certain circumstances he may lash out and release all his aggression in one, often disastrous, act. Afterwards he reverts to his usual overcontrolled defenses. Thus he may be more of a menace than the verbally aggressive "chip-on-the-shoulder" type who releases his aggression in small doses.

This suggests the hypothesis that assaultive criminals can be divided into at least two quite distinct personality types: the Undercontrolled Aggressive type and the Chronically Overcontrolled type.

The Undercontrolled Aggressive person corresponds to the typical conception of an aggressive personality found in the literature. He is a person whose inhibitions against aggressive behavior are quite low. Consequently, he usually responds with aggression whenever he is frustrated or provoked. Since inhibitions are specific to the situation, he will, occasionally, be inhibited from expressing his aggression. For instance, he might not attack his mother or a judge even though they frustrate him. In such cases, however, the Undercontrolled Aggressive person will readily use the mechanism of displacement and find a substitute target for his aggression or he may resort to the mechanism of response generalization and make a less drastic response to the original frustrating agent. Because of his low level of inhibitions he is likely to be diagnosed as a sociopathic personality, antisocial or dyssocial type. Hence his personality dynamics are likely to be similar to those of many other people who have legal difficulties.

The Chronically Overcontrolled type behaves quite differently, however. His inhibitions against the expression of aggression are extremely rigid so he rarely, if ever, responds with aggression no matter how great the provocation. These inhibitions are not focused on a few specific targets, as was the case with the Undercontrolled Aggressive type, but instead are quite general. He is, therefore, unable to make use of the mechanisms of displace-

[3] The writer tends to classify aggression as "extreme," "moderate," or "mild." The term "extreme" is reserved for physical aggression of homicidal intensity; the term "moderate" is used to describe physical aggression less likely to kill or maim the victim and in which there is more adequate justification for the aggressive response; "mild" is a term reserved for most verbal aggression and for physical aggression which is not likely to seriously injure the victim. Most schoolyard scuffles, the majority of "fouls" in sporting events, and such laboratory procedures as administering shock fall into this category. More precise operational definitions of "moderate," and "extreme" assault will be given in the procedures section and in Appendix A.

ment or response generalization. The result is that through some form of temporal summation, such as described by Dollard, Doob, Miller, Mowrer, and Sears (1939, p. 31), his instigation to aggression builds up over time. In some cases, the instigation to aggression summates to the point where it exceeds even his excessive defenses. If this occurs when there are sufficient cues to aggression in the environment, an aggressive act should result.

Because the inhibitions are so excessive, it would appear that when the Chronically Overcontrolled person finally does commit an aggressive act, his instigation to aggression should typically be at a higher level than that of the Undercontrolled or Habitually Aggressive person, simply because more instigation is needed to overcome such excessive inhibitions.[4] If we assume that the degree of violence of the aggressive act is proportional to the degree of instigation, this suggests a way that this typology might be empirically verified. It would follow that a group of people who have committed extremely aggressive acts such as homicide or assault with a deadly weapon would be likely to include some people of the Chronically Overcontrolled type and some of the Undercontrolled Aggressive type. A group of people who have engaged in moderately aggressive behavior, such as fistfights, should on the other hand, consist almost exclusively of the Undercontrolled Aggressive type. On various indexes or measures of aggressiveness and control, then, the extremely assaultive group should appear less aggressive and more controlled *as a group* than would either the moderately aggressive group or a nonassaultive sample. If, on the other hand, the prevailing view is correct and all assaultive people are undercontrolled, then an extremely assaultive group should show the most aggression and the least control relative to other groups.

[4] This is not necessarily the case, of course. If the Undercontrolled Aggressive person has been severely frustrated or provoked, it is possible that his instigation level, too, will be high. But by and large it is likely that most provocations will not be that extreme.

REVIEW OF THE LITERATURE

No prior studies have systematically examined this hypothesis. Nevertheless, there are data in the literature which are relevant to it. The first source of data comes from the literature on psychological tests. In an effort to validate various tests or scales of aggression, a typical procedure has been to administer the test to "nonaggressive" and "aggressive" groups and observe the difference, if any. When the aggressive group is mildly or moderately aggressive, we would expect it to show more aggression and less control than the nonaggressive group. If, however, the aggressive group has engaged in extreme or homicidal aggression, then the typology outlined above would predict a "reversal," that is, a measurement of the extreme group as being less aggressive or hostile and more controlled than the contrast group.

The majority of studies have been of the first type: that is, they have used mild or moderately aggressive people in their criterion group. When significant differences have been found, they show the aggressive group as having higher aggression or hostility scores on the tests. Purcell (1956) gave the Thematic Apperception Test (TAT) to three groups of army trainees referred for psychiatric study and found the most aggressive group to be highest in need Aggression (n Agg). Young (1956) also found high n Agg scores among her sample of institutionalized delinquents, while Mussen and Naylor (1954) found a significant positive relation between the amount of overt aggression displayed in detention and the amount of n Agg in a sample of juvenile delinquents.

Results are similar with the Rorschach. Studies of assaultive hospital patients have shown them to be higher on a number of scales of hostile content (Finney, 1954; Sommer & Sommer, 1958; Storment & Finney, 1953; Towbin, 1959). In a group of convicts, Rader (1957) found a positive correlation between his Rorschach hostile-content scale and aggressive remarks made in group therapy sessions, while Gorlow, Zimet, and Fine (1952) reported that delinquents scored higher than

nondelinquents on Elizur's (1949) scale of hostility.

Thus, studies of mild and moderately assaultive psychiatric patients or delinquents indicate that they have fewer inhibitions against expressing aggression on projective tests than do nonaggressive groups. Among normal subjects also, the usual finding has been that the amount of aggression shown on the test varies directly with a number of criteria of overt aggression (Elizur, 1949; Lindzey & Tejessey, 1956; Murstein, 1956; Pattie, 1954; Walker, 1951).

The first aspect of the typology, therefore, appears to be well established: mildly aggressive and moderately aggressive subjects are relatively undercontrolled. This is, of course, in accord with the prevailing view. The critical question is whether extremely assaultive subjects are overcontrolled. The data here are much less adequate, but a few studies have used subjects who might be classified as extremely assaultive. Stone (1953) administered the TAT and Rorschach to three groups of military prisoners who differed in aggressiveness. The most aggressive group consisted of 31 men in prison for assaults or murders who had at least two prior offenses of this type. Stone got mixed results. He found that on the TAT the most aggressive group manifested significantly more aggression than did the other two. On the Rorschach, however, the medium aggressive group had significantly more aggression than did the most aggressive group, which in turn was tested as significantly more aggressive than the least aggressive group.

While it is difficult to explain these data fully, it is clear that they do not offer much support for the hypothesis that extremely assaultive people are often overcontrolled. However, it should be pointed out that the mixing of assault cases with murderers in the most aggressive group (with unknown proportions of each) would work against the hypothesis. More important, limiting the most aggressive group to men who had at least two prior assaultive offenses would almost certainly screen out Overcontrolled Assaultive offenders and limit the group to the Undercontrolled Assaultive type.

A study by Weinberg (1953) is more relevant to the present writer's hypothesis. He also used three groups, but his test measure of aggressiveness was the Rosenzweig Picture-Frustration (P-F) Study. Weinberg's first group consisted of 22 male prisoners in the Oregon State Penitentiary for felonious assault, assault with a deadly weapon, or assault with intent to kill. This sample can be considered extremely assaultive. His second group consisted of 27 nonassaultive prisoners who were confined for forgery. His third group consisted of 43 normal, noninstitutionalized job applicants who were matched with the prisoners for occupation, age, and education. All subjects were told that the test was for research purposes only.

Weinberg found that the extremely assaultive group obtained Extrapunitiveness scores significantly below those of the forgers, who in turn scored significantly lower than the normals. In a personal communication to Weinberg, Rosenzweig suggested that perhaps the prisoners censored their responses, but Weinberg pointed out that the job applicants were also motivated to appear in a good light. It is clear, however, that prisoners can, and sometimes will, alter their P-F responses so as to appear less aggressive (Megargee, 1964). However, the fact that the extremely assaultive group was significantly lower than the nonaggressive group of prisoners, who were probably equally motivated to dissimulate, is consistent with the hypothesis that extremely assaultive prisoners as a group may be relatively overcontrolled.

A study by Megargee and Mendelsohn (1963) set out to test this hypothesis directly. Three groups of criminals who were candidates for probation were compared on an index based on Murstein's (1956) Rorschach Hostility Scale. The extremely assaultive group consisted of 21 men who had been convicted of murder, assault with a deadly weapon, voluntary manslaughter or mayhem. The moderately assaultive group consisted of 21 men convicted of battery. The nonviolent criminal group consisted of 27 men randomly selected from those convicted of nonaggressive crimes. It was predicted that the

moderately aggressive group would score highest on the Rorschach Hostility Index, the nonviolent group next, and that the extremely assaultive group would score lowest.

As the data in Table 1 indicate, the trend of the data was in the predicted direction; however, the differences between the groups failed to reach acceptable levels of statistical significance when tested by analysis of variance.

A more recent study by these same writers in which an assault scale for the MMPI (*As-3r*) was derived and cross-validated lends additional support to this hypothesis.[5] On the cross-validation of this scale, the extremely assaultive group (convicted of murder, manslaughter, mayhem, or assault with a deadly weapon) scored significantly higher than moderately assaultive, nonviolent criminal, or normal groups. Examination of the content of the scale (which had been derived from empirical item analyses) showed that the items were surprisingly passive and nonaggressive in nature. Moreover, examination of the MMPI validity scales indicated that this was not the result of dissimulation. The *As-3r* scale was found to correlate positively with scales of repression, conformity, and control and negatively with scales of hostility and acting-out such as *Pd*. In short, the pattern that emerged was consistent with what would be expected if the scale were detecting Chronically Overcontrolled people whose repressed hostility had broken through into behavior (Megargee, 1965).

The literature on psychological tests thus supports the notion that mildly or moderately aggressive people show more aggression (and less inhibition of aggression) on psychological tests than do nonaggressive people. The literature on the test performance of extremely aggressive subjects is much less conclusive; however, some findings were noted which were at least consistent with the notion that extremely assaultive people may be relatively overcontrolled.

[5] Megargee, E. I., and Mendelsohn, G. A. The assessment of the chronically overcontrolled assaultive offender. Mimeographed manuscript, 1965. Available on request from the senior author, Psychology Department, University of Texas, Austin, Texas 78712.

TABLE 1
RORSCHACH HOSTILITY INDEXES FOR THREE GROUPS OF CRIMINALS FROM MEGARGEE AND MENDELSOHN (1963)

	N	Mean RHI	*SD*
Extreme assaultive	21	4.134	3.10
Moderate assaultive	21	7.137	10.11
Nonviolent criminals	27	6.197	5.31

A more fruitful, although less rigorous, source of data about extremely assaultive people comes from case studies reported in the literature. Demographic studies consistently show that a large proportion of persons convicted of homicide has no prior history of assaultive behavior (Berg & Fox, 1947; Berkowitz, 1962; Wolfgang, 1957). Moreover, they are generally better behaved while incarcerated and have lower recidivism rates after release than do most other groups of prisoners (Berkowitz, 1962, p. 318). This is, of course, the pattern we would expect if they were Chronically Overcontrolled.

Adolescent murderers have been found to come, for the most part, from adequate or superior homes and to have excellent reputations prior to the offense (Stearns, 1957; Wickham, 1956). In his study of all teenage murderers referred to a court clinic, Wickham (1956) noted that most suffered from a lack of socially acceptable emotional outlets thereby building up tensions and pressures which resulted in a crime of violence.

Schultz (1960) studied four probationers who had assaulted their wives with intent to kill. He found in general, "... a submissive, passive individual who avoided conflict at all costs." He noted a pattern of extreme dependency with rigid control over aggressive impulses as long as the dependency was gratified. When the wife permanently withdrew this gratification by leaving or taking a lover, the control system broke down, and the murderous assault took place.

Lamberti, Blackman, and Weiss (1958)

and Weiss, Lamberti, and Blackman (1960) studied a group of 13 people who, without any prior record of antisocial behavior, suddenly committed a homicide. Their findings were in striking agreement with Schultz (1960). They found the mothers of these murderers had emphasized conformity to the rules of the social system. To gain affection the future murderers had had to deny or repress any hostility. Both clinically and on tests they appeared introverted, insecure, helpless, and unable to assert themselves. The authors concluded, "... [the patients'] difficulties came about because of their needs to conform and because of their inability to act out hostility in ways which they would feel might still be socially acceptable [Weiss, Lamberti, & Blackman, 1960, p. 675]."

Kahn (1959) compared murderers and burglars on a battery of tests and case history data and concluded that the murderers had been significantly more stable and conforming than the burglars. He found the murderers to have personalities which could permit breakthrough of ordinarily rigidly repressed sadistic hostility and to have fewer personality resources for expression of feelings.

These case studies of extremely assaultive subjects are thus consistent with the typology of assaultive offenders which has been suggested. However, a more systematic and rigorous study of the hypothesis is obviously called for. While the case studies cited above represent all those found in a fairly thorough, although by no means exhaustive, search of the psychological literature, sources of bias could easily enter. Clinicians are naturally much less prone to report cases which conform to the general expectation, and editors in turn are less likely to devote journal space to such studies. What is needed, obviously, is a study in which the subjects are selected without bias and in which systematic quantitative observations are made of several groups falling at different points along the continuum of aggressive behavior. The study to be described was designed to meet this need.

SUBJECTS AND GENERAL PROCEDURES

In order to evaluate the hypothesis that extremely assaultive subjects, as a group, will be measured as being low in aggression and high in impulse control, four groups of male juvenile delinquents were selected for study. In the first two groups were all 30 boys who had been detained for serious assaultive crimes in the Alameda County, California, Juvenile Hall during the 10-month period from July 1, 1962, to May 1, 1963.[6] In June 1963, after data collection was completed, the probation officers' reports of the offenses written for the Juvenile Court were collected and examined. The crimes were then rated for amount of aggressiveness on a 10-point Aggression scale devised by the writer. This scale took into account not only the behavior of the defendant, but also such variables as the degree of provocation, the subcultural setting, the immediate stimulus situation, the relative size and armaments of victim and defendant, and the extent of injuries. (See Appendix A.) Ratings were made by the investigator, who had had 3 years experience working with delinquents, and another clinical psychologist with 8 years of such experience. Preliminary ratings were made, serious discrepancies discussed, and the final ratings of aggressiveness made independently. Adequate reliability was achieved with a correlation of .94 between the two sets of final ratings. When discrepancies existed the final ratings for each subject were averaged. (See Appendix B for descriptions of the offenses and the final ratings.)

The scale was then dichotomized and the nine subjects who had scored in the range from 6.0 to 10 were operationally defined as the Extremely Assaultive (EA) group. This group included two cases of homicide, an attempted murder, five assaults with a deadly weapon, and one particularly brutal beating. The remaining 21 subjects, who had scored below 6.0, were defined as being Moderately Assaultive (MA). This group consisted primarily of cases of battery and gang fights.

Since these two groups together comprised all the seriously assaultive delinquents apprehended during this 10-month period, the 30 assaultive subjects were regarded as a population and all such factors as race, age, intelligence, and so forth were left free to vary in accord with the principles of representative design. As it developed the EA subgroup tended to be somewhat younger, have fewer Negro, and more first offenders than the MA subgroup. (The latter relationship was, in fact, predicted.)

In order to add generality to the study and test the hypothesis that EA offenders tend to be overcontrolled relative to other delinquents rather

[6] Only those assaults in which the injury of the victim appeared to be the primary motive were included. Other assaults for other ends during which the victim may have been incidentally injured were excluded. For instance, no cases of forcible rape are included in the sample. (See Appendix B).

TABLE 2

SUBJECT COMPOSITION OF THE FOUR GROUPS

Variable	Group			
	EA	PO	I	MA
N	9	26	20	21
Age range	14–11 to 16–9	11–1 to 17–4	11–2 to 17–9	11–3 to 17–7
$\bar{x}$ Age	14–5	15–5	15–3	15–4
% Negro	44.5	57.7	60.0	66.7
% First detention	77.7	23.0	35.0	28.6
$\bar{x}$ IQ	93.8	91.8	97.3	97.0
IQ range	73–125	67–107	64–140	71–147

than relative only to MA delinquents, two contrast groups of nonassaultive delinquents were also selected for study. They were matched for race, age, and recidivism rate with the total assaultive population (Groups EA and MA combined). The first contrast group (Group I) of 20 boys was selected from among those boys detained for Incorrigibility: unruliness, defiance, and unmanageability in the home. This group was included since it was felt that they were likely to be high on verbal aggressiveness. The second (Group PO) was selected from among those boys detained for property offenses such as auto theft or burglary. (See Table 2.)

Neither of the two contrast groups included any boys who had known records for assaultive crimes. Boys known in advance to be mentally retarded (i.e., to have an IQ below 70) were also excluded from the study.

Each of the 76 subjects was observed during the first 10 days of detention by the custodial staff (which was not informed of the hypotheses being tested) of the unit to which he was assigned. At the end of the third day of detention, each counselor filled out a behavior check list and a set of behavior rating scales.[7] (See Appendixes C and D.) At the end of 10 days, a second behavior check list and set of rating scales was filled out in addition to the Gough Adjective Check List.[8]

During this period, each boy was examined by a clinical psychologist (other than the investigator) from the Probation Department Guidance Clinic. The boy was not told that he was a research subject and was treated like any other diagnostic referral to the clinic with the exception that he was given a standardized interview and test battery and his responses were tape-recorded. As in the case of any referral the boy was told that the psychological assessment was for the purpose of aiding the court in deciding on a disposition. This had the advantage of insuring that the results of the testing could be generalized to the routine clinical situation; it had the disadvantage of encouraging the boys to present themselves in the most favorable light.

The standardized interview was a condensation of that used by Bandura and Walters (1959) in their study of adolescent aggression. The questions used focused on the subject's aggressive behavior toward teachers, parents, and peers. Included in the test battery were the California Psychological Inventory (CPI), the Rosenzweig P-F Study, the TAT, the Holtzman Inkblot Test (HIT), and a brief intelligence measure consisting of the Information and Picture Completion subscales of the Wechsler Intelligence Scale for Children (WISC) or the Wechsler Adult Intelligence Scale (WAIS).[9]

Verbatim typescripts of the recorded interview, TAT, and HIT were prepared by a stenographer, who also removed identifying information from the test protocols and assigned each an identification number from a table of random numbers. The tests and interviews were then turned over to the writer for scoring. The Rosenzweig P-F Study, CPI, intelligence measure, and HIT were scored by standard procedures. In the case of the interviews, ratings were made on the scales prepared by Bandura and Walters (1959), while for the TAT the scoring procedures used by Mussen and Naylor (1954) were adopted.[10]

A final source of data was the Probation Officer's report to the court which contained a social his-

[7] In order to ensure maximum comparability with the Mussen and Naylor (1954) study, the behavior check list and rating scales as well as the directions to the raters, were exact duplicates of those used by them. The writer is grateful to H. Kelley Naylor for providing these forms as well as detailed instructions for scoring the TAT according to his system.

[8] The writer is grateful to Harrison G. Gough for granting limited permission to duplicate his Adjective Check List for use in this study.

[9] The Information subscale was chosen because of all the verbal scales it has the highest correlation with the WAIS Full Scale IQ for 18–19 year olds and with the WISC Full Scale IQ for 13 year olds. The Picture Completion subtest was chosen because of all the performance scales it has the highest correlation with the WISC Full Scale IQ for 13 year olds (Wechsler, 1949, 1955).

[10] Detailed descriptions of the instruments and scoring procedure will be found when each instrument is discussed individually below.

TABLE 3

SUMMARY OF VARIABLES AND HYPOTHESES

Source	No.	Variable name	EA < PO + I + MA	EA > PO + I + MA	EA < MA	EA > MA	EA < I
Predetention behavior							
Probation report	1	Incidence of first offenders				H-1	
	2	Good school attendence		H-2		H-3	
	3	Good school conduct		H-4		H-5	
	4	Incidence of solitary offenses				H-6	
Behavior in detention							
Behavior check list	5	Total verbal aggression	H-7				H-8
	6	Total physical aggression	H-9		H-10		
Rating scale	7	Combined global ratings		H-11		H-12	
Adjective check list	8	Overcontrol adjective index		H-13		H-14	
Psychological examination							
Structured interview	9	Reported physical aggression against peers	H-15		H-16		
	10	Reported physical aggression against authorities	H-17		H-18		
CPI	11	Self-control		H-19		H-20	
Rosenzweig P-F study	12	Extrapunitiveness	H-21		H-22		
TAT	13	Need aggression	H-23		H-24		
	14	Hostility	H-25		H-26		
HIT	15	Movement minus color		H-27		H-28	

tory, a description of the offense and the individual's past criminal record.[11]

A total of 28 specific predictions were made concerning the various dependent variables. All tested aspects of the general hypothesis that the EA group would be lower on measures of aggressiveness and higher on measures of control than the other groups in general and the MA group in particular. In the case of measures of verbal aggressiveness it was hypothesized that Group EA would be lower than the Incorrigible (I) group in particular. Table 3 summarizes the various hypotheses.

The statistical tests used varied as a function of sample size and level of measurement. For Hypotheses 1 through 6 only classificatory or nominal scale data were available so Fischer's Exact Probability Test was used for the small sample comparisons of the EA and MA groups while an adaptation of the binomial test was employed for the larger sample comparisons of the EA group with the rest of the sample. For the remaining hypotheses ordinal scale measurement was attained so the Mann-Whitney *U* Test was employed (Siegel, 1956). For all but two of the hypotheses these tests made it possible to report the exact probability. Since directional predictions were made throughout, all the *p*s reported are one-tailed.

[11] An effort was also made to see all the parents for structured interviews condensed from those used by Bandura and Walters (1959). However, personnel problems, lack of cooperation on the part of some parents, as well as technical difficulties with the recording equipment, so limited the number of usable interviews that the procedure was dropped from the data analysis.

RESULTS

Predetention Behavior

One difficulty with studies using criminal or delinquent subjects is that the commission of the offense and the subsequent judicial procedures may change the person and influence the measures obtained. Therefore, efforts were made to secure data about behavior occurring prior to apprehension. Hypotheses were made about four aspects of behavior typically available for juvenile offenders: the number of prior detentions, the school attendance and conduct records, and whether the actual offense was committed alone or as part of a group.

Hypothesis 1. If the EA group contained a greater proportion of Chronically Overcontrolled people, then it would be expected that this group as a whole would have experienced fewer prior incarcerations in Juvenile Hall than the other delinquent groups. Since Groups I and PO had been selected to match the total assaultive population on this variable, no comparison with them could be made. However, recidivism played no part in the selection of Group MA. Accordingly, it was hypothesized that Group EA would have fewer subjects with prior detentions for any offense than would Group MA. The data in Table 4 support this contention. Only 22% of the Group EA boys had prior detentions, while over 70% of the Group MA boys had been previously confined, a difference significant at the .02 level.

TABLE 4

INCIDENCE OF FIRST OFFENDERS FOR THE EXTREMELY ASSAULTIVE AND MODERATELY ASSAULTIVE GROUPS

	Extremely assaultive	Moderately assaultive	Total
First detention	7	6	13
Recidivist	2	15	17
Total	9	21	30

Note.—$p = .02$.

Hypotheses 2 and 3. It was predicted that the EA subjects would have better school attendance records than the other groups in general and Group MA in particular. Fifty-six of the 76 court reports included information concerning school attendance records. Attendance was categorized as Satisfactory or Unsatisfactory. Subjects who were not attending school due to suspension, exemption, or expulsion were not included in this analysis unless the nature of the attendance prior to the suspension was noted.

In Table 5, the attendance records of the EA subjects are compared with those of the MA group and the rest of the sample. To test the hypothesis that the EA group would have better attendance than the MA group (Hypothesis 3), the Fisher Exact Probability Test was used (Siegel, 1956). This resulted in a p of .084.

For the comparison of EA with the rest of the sample, including MA, an adaptation of the binomial test was employed. The proportion, P, of subjects with Satisfactory attendance records was calculated for the combined PO, I, and MA groups and found to be .34. The binomial test, corrected for continuity, was then used to determine the probability that the proportion of .84 which was observed in the EA sample was a chance deviation within the same population represented by the other samples (Guilford, 1956, p. 175 ff.; Siegel, 1956, pp.

TABLE 5

SCHOOL ATTENDANCE RECORDS OF THE FOUR GROUPS WITH p VALUES OF THE TESTED DIFFERENCES

	Group				Total	p value of tested comparisons	
	EA	PO	I	MA		EA versus PO, I and MA	EA versus MA
Satisfactory	6	5	4	7	22	.003	.084
Unsatisfactory	1	14	10	9	34		
Total	7	19	14	16	56		

TABLE 6

SCHOOL-CONDUCT RECORDS OF THE FOUR GROUPS WITH p VALUES OF THE TESTED DIFFERENCES

	Group				Total	p value of tested comparisons	
	EA	PO	I	MA		EA versus PO, I and MA	EA versus MA
Satisfactory	3	4	2	8	17	.166	.258
Unsatisfactory	3	14	11	12	40		
Total	6	18	13	20	57		

36–42). This test resulted in a z of 2.56 which had a one-tailed p of .003. As Siegel (1956) has pointed out, this procedure is not usually recommended with samples as small as this; however, the p value is so small it is unlikely that we would commit a Type 1 error in rejecting the null hypothesis.

Hypotheses 4 and 5. It was also hypothesized that EA subjects would have better school-conduct records than the other subjects. Reports were available for 57 subjects and, as in the case of attendance, the conduct reports were classified as Satisfactory and Unsatisfactory. Data were available on six of the EA group; 50% had Satisfactory ratings as compared with 22% of the PO group, 15% of the I group, and 40% of the MA group. The difference between the EA and MA groups had a p of .258 when tested with the Fisher Exact Probability Test, while the difference between Group EA and the rest of the sample had a p of .166 when tested with the binomial test.

Hypothesis 6. For the EA group, aggression was assumed to be ego alien rather than ego syntonic. If so, the assault would be apt to be a furtive act committed while alone rather than a socially acceptable act committed while with others. Accordingly, it was hypothesized that the EA group would have a greater proportion of offenders in which the defendant and victim were alone at the time of the offense than would the MA group. In Table 7 the data are presented.

TABLE 7

INCIDENCE OF SOLITARY VERSUS GROUP OFFENDERS FOR EXTREME ASSAULTIVE AND MODERATE ASSAULTIVE GROUPS

	Extreme assaultive	Moderate assaultive	Total
Alone	6	4	10
Group	3	17	20
Total	9	21	30

Note.—p = .021.

The data show that while two thirds of the EA offenders were alone with their victim, less than 20% of the MA subjects were. The Fisher Exact Probability Test resulted in a p of .021. This finding is interpreted as indicating that physical aggression is more socially acceptable for the MA subjects. [An alternative explanation would be that the MA subjects are simply more outgoing and friendly than the EA subjects. This hypothesis, however, is contraindicated by the ratings scales to be described below (Hypotheses 11 and 12), on which it was found that, on the contrary, the EA subjects were rated as being significantly more cooperative (p = .01) amiable (p = .051), and friendly (p = .02) than the MA subjects when their social interactions were observed during the first 10 days of custody.]

Summary of Predetention Data. Six predictions were made regarding behavior occurring prior to any judicial action. All the relationships were in the predicted direction, and three were highly significant while another attained marginal significance. These data indicate that even before coming into custody the EA boys behaved in a manner consistent with the notion that they are overcontrolled and inhibited in the

expression of antisocial tendencies relative to other groups of delinquents.[12]

BEHAVIOR IN DETENTION

While in detention awaiting court hearings, the subjects, like all other boys in Juvenile Hall, were assigned to one of four custodial units. Each unit contained approximately 40 boys and was supervised from 7:30 A.M. to 11:30 P.M. by two sets of two counselors working 8-hour shifts. These men were with the boys constantly during the daylight hours and as a routine matter observed each boy's behavior and interactions during recreation periods, sports activities, meals, and work assignments. For the boys included in the study, each counselor filled out a behavior check list and a set of rating scales on the third and tenth day of the boy's detention and a Gough Adjective Check List on the tenth day. These first two instruments are duplicates of those used by Mussen and Naylor (1954) and are reproduced in Appendixes C and D.

The behavior check list, originally devised by Naylor (1952), listed 13 categories of aggressive behavior (See Appendix C). Each counselor in the unit checked off each category of aggressive behavior he had observed each subject engage in. Because of days off, vacations, and sick leave, the counselor population was rather fluid, resulting in anywhere from 7 to 14 individuals coming in contact with each boy. The number of reports received on each subject varied accordingly, so the number of reports listing a specific category of behavior for a subject was divided by the total numbers of reports submitted on that subject and multiplied by 100 yielding a percentage score. Thus, if a boy had nine behavior check lists submitted on him, and three listed Physical Attack, his score on this variable was 33.3.

[12] These data might also be interpreted as reflecting extreme undercontrol by the MA group rather than overcontrol by the EA group. The Adjective Check List and Holtzman Inkblot Technique data reported below (Tables 8 and 9) contraindicate this interpretation however, as do the significant differences found when the EA group was compared with the other three groups combined.

Seven of the categories on the behavior check list (Bragging; Teasing; Saucy, Impertinent; Insulting, Name Calling; Ridiculing, Mocking; Verbal Castigation; and Malicious Gossip) seemed to reflect verbal aggression so the percentage scores for these categories were added to give a score for Total Verbal Aggression. In like manner, the scores for five categories (Physical Attack, Threatening, Bullying, Destructive, and Temper Tantrums) seemed to reflect physical aggressiveness, and these were combined into a score for Total Physical Aggression.[13]

Hypotheses 7 and 8. It was hypothesized that EA subjects would be lower than the combined contrast groups on Total Verbal Aggression (Hypothesis 7) and that the EA subjects in particular would be lower than the group (Hypothesis 8) which was expected to be high in verbal aggressiveness. The data are presented in Table 8. As expected, Group I had the most Verbal Aggressiveness and Group EA the least.

When Hypothesis 7 was evaluated by means of the Mann-Whitney U Test, an exact probability of .058 was obtained.[14] In the analysis of Hypothesis 8, the EA group was contrasted with Group I and the difference found to be significant with $p < .05$.

Hypotheses 9 and 10. It was predicted that EA subjects would be the lowest on Total Physical Aggression on the behavior check list (Hypothesis 9) and, moreover,

[13] The thirteenth category, used by Mussen and Naylor (1954), "Running Away," did not appear to fit in either of these categories, and indeed appears to be fairly nonaggressive in nature. Therefore, it was not included in the analysis.

[14] In this analysis, as in most of those to be reported, the EA subjects were contrasted with the other three groups combined. This was done because, for these hypotheses, the writer was always predicting that Group EA would be higher or lower than the other groups; other differences obtained between the means of the other groups were consequently not relevant to the primary concern of the study. The reader may also have noted that the total number of subjects in this analysis is 75. Inevitably, with this many variables and sources of data, there was some missing information for practically every variable. In this case, one boy was released by the court after his testing and interview were completed, but before observational data could be collected.

TABLE 8

MEAN SCORES OF THE FOUR GROUPS ON MEASURES OF BEHAVIOR IN DETENTION WITH p VALUES OF THE TESTED DIFFERENCES

Variable		Group scores				p values		
		EA	PO	I	MA	EA versus PO, I and MA	EA versus MA	EA versus I
Behavior check list								
Total verbal aggression	M	74.41	113.44	137.43	103.80	.058		<.05
	SD	79.04	88.52	92.85	73.07			
	N	9	26	19	21			
Total physical aggression	M	27.04	32.78	44.17	51.02	.374	.255	
	SD	25.84	50.26	49.23	58.31			
	N	9	26	19	21			
Combined rating scales	M	16.11	15.36	14.96	14.80	.06	.045	
	SD	1.97	1.98	1.52	1.77			
	N	9	26	20	21			
Adjective check list								
Overcontrol index	M	2.42	0.56	−0.78	−0.57	.028	.055	
	SD	3.07	3.98	3.29	4.15			
	N	9	26	20	19			

that in particular Group EA would be lower than Group MA (Hypothesis 10). While the differences were in the expected direction, they were far from significant.

It should be remembered, however, that these ratings were made in a custodial setting. Not only was swift punishment administered for any physical aggression, but also the boy knew that his behavior in detention would undoubtedly influence the court disposition. Such external controls would tend to reduce the amount of aggression engaged in by any subjects of the Undercontrolled Aggressive type and hence work against the hypothesis.

Rating Scales

Hypotheses 11 and 12. The second measure of overt behavior during detention was the set of 5-point rating scales devised by Naylor (1952) for a study of juvenile delinquents. The set consisted of five bipolar scales ranging from an unfavorable trait such as Uncooperative or Aggressive to a more passive or favorable one such as Cooperative or Submissive. Each scale was anchored at each point by a brief description of the behavior which would earn such a rating. (See Appendix D.)

These five rating scales were combined into a global scale with a possible range from 5 to 25. High scores reflected cooperativeness, amiability, submission, docility, and friendliness while low scores indicated an uncooperative, quarrelsome, aggressive, rebellious, and antagonistic attitude. It was hypothesized that the EA group would have a higher score than the other groups combined (Hypothesis 11) and the MA group in particular (Hypothesis 12). This expectation was confirmed with the EA group having the highest score and the MA group the lowest. Hypothesis 11 had a p of .06 while Hypothesis 12 had a p of .045. (See Table 8.)[15]

Gough Adjective Check List

Hypotheses 13 and 14. On the tenth day of detention, each counselor checked all those adjectives on the 300-item Gough Adjective Check List which he felt were descriptive of the boy. Forty of the adjectives were selected for study. Twenty of these were adjectives which seemed de-

[15] Not only was the EA group rated most favorably on the combined scales, but it also had the most favorable rating on each of the individual scales as well.

scriptive of the Chronically Overcontrolled person such as "meek," "self-controlled," "conscientious," and "withdrawn." Twenty others seemed descriptive of the Undercontrolled Aggressive type, including such terms as "aggressive," "hostile," "irritable," and "assertive." (See Appendix E.) The adjective check list submitted by each counselor was scored by counting the number of adjectives of each type. Then an Overcontrol index was created by subtracting the Undercontrolled Aggressive adjectives from the Chronically Overcontrolled ones. It was hypothesized that the EA group would have the highest score on this index.

The data were in the predicted direction with a p of .028 when Group EA was compared with the combined contrast groups and .055 when Group EA was compared with Group MA. (See Table 8.)

The Overcontrol index was particularly noteworthy in another regard. The three contrast groups, MA, PO, and I, all had quite similar scores as would be expected if they shared essentially the same values and orientation. The EA group, on the other hand, had a score almost five times that of the next highest contrast group. This is consistent with the basic thesis that the EA offender category is apt to include a distinctly different type of person than other offenders. Secondly, this difference is shown by the Overcontrol index data to be clearly in the direction of excessive control on the part of the EA subject rather than extraordinary aggressiveness on the part of the MA subject.

Summary of Behavior in Detention

On all of the devices used to assess behavior during detention, the EA boys were measured as being less aggressive and more controlled than were the members of the other three groups. The consistency of these results adds confidence to the reliability of these observations.

A logical question is whether or not this reliability was the product of some set or unconscious bias among the observers. In this regard it is important to recall that the counselors who made the ratings had no idea what hypotheses were being tested, having been told merely that it was a study on aggressiveness. Knowing this and knowing the offenses with which each boy was charged, one would expect, if anything, a set to rate the boys charged with extremely assaultive crimes as being more aggressive. If such a set did exist it would operate against the actual hypotheses, so that the differences obtained were in spite of such a set rather than because of it.

Secondly, the obtained differences were also in spite of the fact that the boys were confined in a custodial setting in which swift sanctions were levied against any aggressive behavior. The effect of this would be to curb the aggressiveness of the Undercontrolled Aggressive boys by providing external controls in place of their deficient internal controls. It should have little effect on Chronically Overcontrolled people. Thus, the setting operated to reduce differences, and it is likely that if the observations could somehow have been made in the natural milieu the differences might have been even more striking.

RESULTS OF THE PSYCHOLOGICAL ASSESSMENT

Structured Interview Data

In order to obtain information concerning the subjects' attitudes toward aggression and their customary behavior "on the street" (i.e., when not in custody) each boy was interviewed. The structured interviews used by Bandura and Walters (1959) were selected for this purpose because reliable scales had been devised by those authors so that responses could be quantified. Time limitations precluded administering the entire schedule, so the interview consisted of the following questions: 1, 3, 4, 5, 6, 9, 10, 13, 22, 31, 33, 34, and 38 (Bandura & Walters, 1959, Appendix B). For the most part these questions asked the subject how he behaved in various situations (i.e., "How do you deal with the kind of guy who likes to push his weight around?") or about the amount of aggressive behavior he had engaged in in the past (i.e., "How often have you gotten into a fight since you've been at high school?").

The interviews were scored on several of

the scales of physical aggression devised by Bandura and Walters (1959): physical aggression against peers, against teachers, against mother, and against father. A single scale of "physical aggression against authorities" was devised by computing the mean for the latter three scales.

Hypotheses 15 through 18. It was hypothesized that the EA group would receive the lowest scores on these two scales of physical aggression. This prediction was not upheld in the case of physical aggression against peers (Hypotheses 15 and 16). On this variable the Property Offenders reported they had engaged in the least aggression; while the score for the EA group was somewhat lower than that for the MA group, the difference was far from significant.

In the case of reported aggression against authorities (Hypotheses 17 and 18) the EA group did receive the lowest score. The p value for the difference between the EA group and the other groups combined was .065, while the difference between the EA and MA groups had a p of .082.

California Psychological Inventory

It was impossible to administer the CPI to every subject since many were unable to read it adequately. In a few cases the test was read to the subject, but limitations on staff time precluded this as a standard operating procedure. After scoring, all CPIs on which the Communality score was less than 20 were discarded as invalid on the basis of probable random answering (Gough, 1960, p. 20). This left a total of only 46 valid CPI protocols from the total sample of 76. The bias introduced by this loss cannot be determined; however the proportion of usable profiles was quite similar for the four groups ranging from a low of 58% for the PO group to a high of 67% for the EA group. The probable effect was to eliminate the least cooperative and least intelligent subjects from each group.

Since the scales on the CPI are such that higher scores reflect more positive traits, it was generally expected that the EA group would score highest on most of the scales. This expectation was upheld, since the EA group had the highest mean score on 13 of the 18 scales.[16]

Hypotheses 19 and 20. It was hypothesized that the EA group would be highest on the Self Control (Sc) scale and this was found to be the case. The significance of the difference between Group EA and the other groups had a p of .129. The exact probability of the difference between Groups EA and MA could not be determined because of the small number of subjects involved, but the Mann-Whitney U value of 30.5 which was obtained was far above the value of 19 required for significance at the .05 level.

It is noteworthy that the mean Sc score of 26.5 obtained by the EA subjects was somewhat above the usual high school norms. This is what would be expected if the EA group contained some subjects who were *over*controlled and not just more controlled than the other delinquents.

Other CPI Relationships. In examining the means on the other scales, it was noted that the EA group scored markedly higher on the Responsibility (Re), Well Being (Wb), Tolerance (To), Achievement by Independence (Ai), Intellectual Efficiency (Ie), and Flexibility (Fx) scales. If these differences had been anticipated and directional predictions had been made, some of them would have been significant. The one-tailed p for the difference between Group EA and the other groups on the Re scale, if it had been predicted, would have been .12, for the Wb scale .01, for the To scale .07, for the Ai scale .07, for the Ie scale .01, and for the Fx scale .06.

This score pattern indicates that the members of the EA group tend to be more conscientious, responsible, and alert to ethical or moral issues than the members of the other groups (Re). They are particularly oriented toward doing well in school and tend to be more mature and thorough in their approach to academic tasks (Ai and Ie). They tend to be more alert, ambitious, and enterprising and are more likely to value work and effort for their

[16] Because of lack of independence among the CPI scales, no statistical test of this prediction was possible.

own sakes (Wb). They appear to be more verbal, tolerant, and clear thinking (To), although they can be sarcastic and cynical in their verbal behavior (Fx), and, of course, as predicted, they are less impulsive and more controlled (Sc). All in all, they tend to have more of the traits which are valued among the middle-class members of our society. This pattern is consistent with the personality pattern hypothesized for the Overcontrolled type as opposed to the Undercontrolled type.

Rosenzweig Picture-Frustration Study

The Rosenzweig P-F Study has been used in a number of studies of aggression (e.g., Weinberg, 1953) and was therefore included in the present investigation. Since it is a relatively unsubtle instrument when administered in the context of a court clinic, there were some reservations about whether or not the results might be influenced by dissimulation on the part of the subjects.

Hypotheses 21 and 22. It was hypothesized that the EA group would be the lowest of the four on the Extrapunitiveness scale of the P-F. This did not prove to be the case. The EA group did have a lower score than the MA group as had been predicted (Hypothesis 22), but the high p value indicated that this was only a chance relation.

In order to determine the validity of these results, the mean scores of the four groups were compared with the normative scores for boys aged 14–19 reported by Deming (1960). It was found that the mean scores for all four of the delinquent groups fell below the mean score of 46.4 reported by Deming. This suggested that dissimulation was influencing the results.

In order to investigate this notion further, the Extrapunitiveness scores for the 46 boys for whom valid CPIs were available were correlated with the CPI Good Impression scale, which was designed to detect "faking good." A correlation of −.46 was obtained which was highly significant ($p < .01$). These data would indicate that in a court setting such as this, the P-F results can be markedly influenced by defensiveness (Megargee, 1964). It is questionable, therefore, how adequately the P-F tested the basic hypotheses.

Thematic Apperception Test

Hypotheses 23 and 24. The TAT was administered in the normal fashion with the exception that the stories were tape-recorded rather than written down by the examiner. An effort was made to employ the same methods used by Mussen and Naylor (1954) in their study of the relation between fantasy and overt aggression. The cards used (1, 3BM, 4, 6BM, 7BM, 8BM, 12M, 13B, 14, and 18BM) were the same, and the cards were scored for n Agg in the same manner using verbatim typescripts prepared from the tape recordings.

The hypothesis that the EA group would be lowest on n Agg was not supported by the data. In fact, Group MA proved to have the lowest scores rather than Group EA.

Holtzman Inkblot Test

The Holtzman Inkblot Test (HIT) was administered to all the subjects using standard procedure. All responses were tape-recorded, and scoring was done by the investigator from verbatim typescripts. The HIT was selected rather than the Rorschach because its scoring procedure is more amenable to statistical manipulation and because the use of 45 cards with one response to each was apt to elicit a larger body of responses. Both content and determinant scores are obtainable from the test, and both were used.

Hypotheses 25 and 26. It was hypothesized that the EA group would be lowest on the Hostility scale (Hs) and that Group EA would be significantly lower than Group MA. This is a content scale based on the one devised by Murstein (1956) for the Rorschach. The data in Table 9 show that on the contrary, the EA group had the highest Hs score, so both hypotheses were disconfirmed. (The significance of this reversal was tested and found to be insignificant, with a two-tailed p of .267.)

Hypotheses 27 and 28. The last hypotheses to be tested were related to the determinants of Movement and Color. In the

Klopfer system, Movement responses are interpreted as indicating "an inner system of conscious values of one kind or another, in terms of which the person tends to control his behavior, to guide his satisfactions, and to postpone his gratifications [Klopfer, Ainsworth, Klopfer, & Holt, 1954, p. 262]." If Movement responses on the HIT have a similar meaning, it would be expected that the EA group would be highest on this variable.

The use of color, on the other hand, is associated with immature and impulsive behavior, particularly when the color is used as a primary determinant with little attention paid to the form elements of the blot. Since the HIT Color score assigns the heaviest scoring weights to these uncontrolled color responses (Holtzman, 1961) it would be expected that the EA group would have the lowest Color scores.

Since the Overcontrolled subjects should be relatively high on Movement and low on Color a simple index was used in which each subject's Color score was subtracted from his Movement score. It was predicted that the EA group would have the highest score on this Movement-Color index.

This prediction was borne out by the data. The Movement-Color score of the EA group was markedly higher than those of the other three groups. The p value for

TABLE 9

RESULTS OF DATA COLLECTED DURING THE PSYCHOLOGICAL ASSESSMENT

Variable		Group scores					p value of tested comparisons	
Instrument	Variable name		EA	PO	I	MA	EA versus PO, I and MA	EA versus MA
Structured interview	Reported physical aggression against peers	*M*	3.39	3.11	3.45	3.50	Not tested[a]	.319
		SD	1.24	1.32	1.31	1.16		
		N	9	23	20	20		
	Reported physical aggression against authorities	*M*	1.08	1.20	1.34	1.19	.065	.082
		SD	.22	.26	.36	.30		
		N	9	23	20	20		
CPI	Self-control	*M*	26.5	22.5	21.8	23.4	.129	$>.05$[b]
		SD	6.13	6.42	6.68	10.14		
		N	6	15	12	13		
Rosenzweig P-F study	Extrapunitiveness	*M*	40.6	39.71	35.46	42.24	Not tested[a]	.48
		SD	7.73	13.91	15.61	19.75		
		N	9	26	20	18		
TAT	Need aggression	*M*	8.33	8.52	10.05	8.10	Not tested[a]	Not tested[a]
		SD	2.79	3.50	5.08	4.52		
		N	9	23	20	19		
	Hostility	*M*	10.33	6.84	8.80	7.86	.267[c]	Not tested[a]
		SD	6.00	5.02	6.10	5.37		
		N	9	26	19	21		
HIT	Movement-color index	*M*	14.00	8.08	5.50	8.18	.061	.059
		SD	8.79	14.91	23.99	12.94		
		N	9	26	19	21		
	Number of pure color responses	*M*	1.11	1.76	2.85	1.86	.111	.045
		SD	1.59	1.97	5.97	1.46		
		N	9	26	19	21		

[a] Data not in hypothesized direction.
[b] Exact probability test not possible because of small N's.
[c] Two-tailed test used to test reversal.

the difference between the EA group and the rest of the sample combined was .061, while the p value for the comparison between the EA and MA groups was .059. This pattern of scores, which resembles that found for the Adjective Check List Overcontrol index, lends further support to the notion that while the behavior and dynamics of the MA group are similar to those of other offenders those of the EA group are qualitatively different as would be the case if a different personality type was involved.

In order to determine whether the relatively greater use of Color by the other three groups was the result of the use of uncontrolled color responses as opposed to a large number of well-controlled color responses, the incidence of pure color responses in all four groups was calculated. The data in Table 9 indicate that, as expected, Group EA had the lowest number of pure color responses. The p of the difference between the EA group and the rest of the sample was .111, while the p of the difference between the EA and MA groups was .045.

Summary of the Results of the Psychological Assessment

The results of the psychological assessment are not as clear-cut as were those of the preoffense behavior or the behavior while in detention. No support was obtained from the Rosenzweig P-F Study, the TAT n Agg measure and the HIT Hostility scale. The interviews indicated that the EA group was less aggressive to authorities but the differences in the amount of aggression against peers reported were not significant. The CPI data were in the predicted direction but with marginal p values. The best support came from the Movement-Color index on the HIT, on which the EA group displayed substantially more impulse control.

It is not too surprising that the behavior on the psychological tests is not as clear-cut as that observed in detention or found in the case history. Studies such as those of Kostlan (1954) and Little and Shneidman (1959) have demonstrated the greater validity of case history data as opposed to psychological tests. This tendency to find greater clarity in direct measures as opposed to tests would probably be accentuated in a correctional setting such as the one in which these data were collected. A delinquent being assessed to aid in determining a court disposition is naturally going to be quite guarded and defensive during a psychological examination, but is less likely to be able to maintain a subterfuge over 10 days of interaction with other delinquents.

Within the psychological test data, the more obvious the instrument, the more likely it is that a defensive attitude could alter the results. This is consistent with the fact that the most obvious tests, the P-F study, the TAT, and the Hostility scale of the HIT failed to show the predicted patterns, but the less easily distorted measures such as the empirically derived CPI Self-Control scale and the Movement-Color index of the inkblot test did show the hypothesized patterns.

DISCUSSION

The first issue to be discussed is whether or not the data presented above support the writer's hypotheses. It will be recalled that the basic hypothesis was that there are two personality types involved in antisocial aggression: the Undercontrolled Aggressive type and the Chronically Overcontrolled type. The former may commit aggressive responses of any intensity depending upon the immediate stimulus situation, while the latter tends to inhibit aggressive responses until they break through in what the writer has called an "extremely assaultive response" in which the very life of the victim may be jeopardized. It followed from this hypothesis that a group of extremely assaultive subjects would be assessed as less aggressive and more controlled, as a group, than would contrast groups of moderately assaultive and other nonassaultive delinquents because of the probable presence of Overcontrolled subjects among the extremely assaultive group, while the latter groups would be made up of the Undercontrolled type.

In the study which was conducted to test this, the results were by no means un-

equivocal in their support for the hypothesis. Nevertheless, by and large, a review of the data indicates consistent if not spectacular support for the writer's hypotheses.

Of the 28 hypotheses, 22 were in the predicted direction with the EA group displaying less aggression or more control than the other groups in general and the MA group in particular. Fourteen of these hypotheses received some measure of statistical support, with p values ranging from .003 to .084. On only 1 of the 15 variables was the EA group assessed as having the most hostility. This difference when tested did not approach significance.

Because of the somewhat marginal significance levels, we cannot consider the case firmly established or definitely proven. However, the total pattern of the data supports the proposed typology and certainly gives no support to the most prevalent opposing notion that *all* extremely assaultive delinquents are *more* aggressive and *less* controlled than other delinquents.

However, as in the case of any experiment, it is possible to advance other ad hoc explanations. For instance, one might argue that the EA group, facing severer penalties for their offenses, behaved in a more controlled fashion during detention in order to impress the court and receive lesser penalties.[17] This hypothesis, however, is not supported by the data. If this were the case, then the EA group would show a pattern of greater control after the offense but not before. However, the predetention measures also showed significantly more social conformity on the part of the EA group in the form of a lower recidivism rate and a better school attendance record.

It is also unlikely that temporary situational constraint could cause significant differences on the Movement-Color index of the HIT. Moreover, undue concern over making a good impression, or even outright dissimulation, probably would be reflected in the CPI Good Impression scale. Yet the EA group's mean score on this scale was almost identical to that obtained by Gough's (1960) normative sample of 3,572 high school males. It would therefore appear that the notion that the results were caused by temporary inhibitions resulting from the judicial process are not consistent with the data.

Another possible explanation of the results is that the EA group appeared less aggressive because they had vented all their hostility during the commission of the offense. This would also account for the fact that all four groups obtained below-average scores on the Rosenzweig P-F Study, since the members of all four groups had probably released more aggressive tensions just prior to testing than had the normative high school sample cited by Deming (1960).

There are two types of data in the study which would contraindicate this drive-reduction hypothesis. The first is the pre-offense behavior cited above. The second is the set of hypotheses in which the EA and MA groups were compared. While it might be argued plausibly that the EA group had undergone more drive reduction as a result of their offenses than other delinquents in general, it would hardly appear likely that the differences in amount of drive reduction between an assault rated as "extreme" and one rated as "moderate" would be sufficient to account for the consistently less aggressive scores of the EA as opposed to the MA group. (See Appendix B.)

Another set of ad hoc hypotheses can be derived from the effects of the representative design used in the selection of the assaultive sample. It would be recalled that all the assaultive delinquents apprehended over a 10-month period were included in the study and that later they

---

[17] The assumption that the EA subjects faced severer penalties at the hands of the Juvenile Court is itself false. The dispositions for juvenile offenders can be dichotomized as (*a*) some form of institutionalization for an indefinite period or (*b*) probation in the community. Of the EA sample, 67% were institutionalized, of the MA sample 63%, of the PO sample 60%, and of the I sample 40%. Thus, while it appears that the Incorrigible subjects were less likely to be institutionalized than were the others, it does not appear that the penalties assigned the EA group were much greater than those meted the rest of the sample. Nevertheless, while the assumption of greater penalties is false, the delinquents themselves could have acted as if it were true.

were subdivided on the basis of the Aggression scale into the EA and MA subgroups. In accord with the basic principles of representative design, no arbitrary restrictions were placed on the subdivision of this population. Consequently the EA group had a greater percentage of white subjects than Group MA or Groups I and PO which were matched to the total assaultive population on this variable. (See Table 1.)

It could be argued, therefore, that the obtained data were the result of the differences in racial balance among the four groups. This would be particularly likely to influence the ratings of detention behavior if some form of stereotype, prejudice, or halo effect were operating in the raters, although it should be pointed out that at least half of the raters were themselves Negro. In order to evaluate this possibility, the detention data were recalculated separately for whites and Negroes. For both white and Negro subjects, the EA group was found to be assessed as least aggressive and most controlled on the measures of behavior while in detention. With the reduced *N*s, the *p* values were of course much higher than they had been for the total sample.

Another ad hoc hypothesis could focus on the differences in recidivism rate over the four groups. (It will, of course, be recalled that this difference was predicted in Hypothesis 1.) However, it could be argued that the differences reported in detention behavior were the result of a negative halo effect for the recidivists on the part of the counselors who were, of course, aware of the past records. In order to evaluate this, the seven first offenders in the EA group were compared with the first offenders in the other three groups on the measures of detention behavior. Even when the study was limited to first offenders, the same directional differences were noted in the detention reports. Once again, the reduced *N*s increased the *p* values.

It would therefore appear that if the prevalent assumption that assaultive criminals are all undercontrolled and highly aggressive is to be maintained, it would be necessary to account for the data obtained by resorting to various ad hoc explanations. However, the most plausible of these ad hoc explanations have been examined, and data have been adduced to demonstrate their inadequacy.

There is one final interpretation of these data which would allow a person to preserve the simple notion that only one personality type is involved in assaultive offenses. This would be to dismiss the present data as simply chance phenomena which do not need explanation. It is, of course, up to the individual to decide how much data he will require before he will abandon the null hypothesis. It is also true that it can be hazardous to rely on directional data and marginal *p* values. Nevertheless, the fact that the present hypothesis predicted results directly opposite those implied by the commonly accepted alternative theory, that these results were quite consistent over a wide range of dependent variables ranging from recidivism rate before the offense to Movement-Color index on the HIT after the offense, and that the significance levels obtained were achieved despite the crudity of criminal offense as an independent variable, all combine to suggest strongly that if the null hypothesis is not rejected, one would be in serious danger of committing a Type 2 error.

The present investigator, as might be anticipated, is more inclined to risk a Type 1 error by rejecting the null hypothesis and accepting the notion that two types of people may be involved in assaultive offenses. If this position is adopted, one implication is that prevailing conceptions of aggression are not always applicable to the dynamics of the extremely aggressive person. It would appear that extreme aggression is a phenomenon which should be studied in its own right and not through extrapolation from studies of milder forms of aggression. As is obvious from the present study, this type of investigation presents many methodological difficulties. The fact that such investigations must take place in a judicial setting not only limits the procedures that can be used without upsetting institutional routines but also will inevitably influence the motivation and set of the subjects. Moreover, the ever-

present psychological problem of the adequacy of our measuring instruments is quite obvious when attempts are made to differentiate levels of hostility within a delinquent or criminal sample (Megargee, 1964; Megargee & Mendelsohn, 1962). Despite these difficulties, however, the present study indicates that if extreme aggressive behavior is to be understood, these problems will have to be coped with since the study of milder aggressive behavior is apt to be misleading.

The study also suggests certain clinical problems. The first is in the area of predicting assaultive behavior. It would appear that there is little difficulty in diagnosing or predicting the behavior of the Undercontrolled Aggressive type. His whole life style should show a pattern of recurring aggression and violence, and there is little doubt that without some dramatic intervention, this style will continue.

The Chronically Overcontrolled type presents a much more difficult problem, however. In the first place, lay personnel tend to overlook the potential pathology of the quiet, retiring person, so that they are much less often referred for evaluation by parents, clergy, or teachers than the Undercontrolled Aggressive type.

Even if an Overcontrolled person is referred, the clinician must somehow discriminate between the Overcontrolled patient who is potentially assaultive and the one who is not dangerous. This may be nearly impossible for it can depend a great deal upon environmental events and frustrations which the clinician can not anticipate.

However, when assaultive people are examined after the offense, there are some indications, in retrospect, that certain cues may have been present which would have indicated potential violence. One is a preoccupation with violence in fantasy. An 11-year-old boy who fatally stabbed his brother was a cartoonist for his school paper, and after the offense people suddenly recalled a cartoon in which his chief character was taking a fencing lesson and stabbed his instructor to death. The boy who obtained the highest rating on the Aggression scale in the present study had shot at his parents from ambush, killing his mother. (See Case No. 07009, Appendix B.) Several months earlier he had thought of writing a novel about a boy who became so disgusted with his parents that he just killed them. Despite these indications of a preoccupation with aggression in fantasy, the apperceptive measures used in the present study generally showed no difference in aggressive ideation among the four groups. Of course these tests were administered in custody after the offense, and a different pattern might have been elicited at some other point in time.

There are some indications in the present study, as well as from other data, that the distinctive pattern that distinguishes the potentially assaultive Overcontrolled person is outward conformity coupled with inner alienation (Megargee, 1965). For instance, despite the docile, controlled pattern displayed by the EA sample in the present study, their Socialization scores on the CPI were in the delinquent range and no different from those of the other delinquent samples. While it will be recalled that the CPI data were biased because of a high percentage of invalid profiles, nevertheless this could indicate that the Chronically Overcontrolled assaultive person shares the typical delinquent's feelings of futility, disgust, and alienation, but instead of acting out these feelings he customarily rigidly represses them. Data collected in connection with the development of an MMPI scale designed to detect the Overcontrolled assaultive offender also are consistent with this pattern. (See Footnote 5, above).

Whether we identify the assaultive person before or after the offense, the question which naturally arises is how he can best be treated so as to make him less dangerous to others. Early identification not only has the obvious advantage of possibly preventing an assaultive act, but also allows greater freedom to choose an appropriate form of therapy. After an offense occurs, legal considerations and public opinion greatly restrict the range of possible choices.

In the case of the Undercontrolled Aggressive type, the basic therapeutic task is

to increase the inhibitions against aggressive acting out. Normally such inhibitions are acquired through identification with a well-socialized parent figure with consequent introjection of his values. However, in the case of the Undercontrolled Aggressive person this has not taken place. If he is treated early enough, it might be possible to foster the growth of such controls by providing a parent substitute in the form of a case worker, clergyman, "big brother," or probation officer. Often, however, this is not feasible, so an alternative program must be used. This usually consists of providing external controls with automatic rewards for approved behavior and punishments for disapproved behavior. In order to control the schedules of reinforcement and protect society during the learning process, institutionalization is generally indicated. Such an institution may be called a camp, a school, a jail, or a penitentiary, but the basic philosophy and the basic program are usually the same.

Unfortunately, such programs are less effective than might be desired. It is difficult, even in an institutional setting optimally to schedule rewards and punishments, with the result that most inmates are on a partial-reward schedule when it comes to the expression of aggression. Instead of learning to inhibit aggression, they are more likely to form a discrimination and inhibit aggression only when they are more likely to form a discrimination and inhibit aggression only when they are likely to be caught. Moreover, the frustrations of life in an institution, as well as the life of an ex-convict, are likely to increase the instigation to aggression enough to offset any increase in inhibitions.

The most appropriate treatment for the Chronically Overcontrolled assaultive person, on the other hand, would be some form of psychotherapy. The goal of such therapy would be to reduce excessive inhibitions so that the individual can learn to acknowledge and accept his feelings of hostility and learn ways of expressing them which would allow some measure of need satisfaction while still not posing too great a threat to society.

If the potentially assaultive overcontrolled person is detected prior to an aggressive outburst, such a treatment program can be instituted fairly easily. However, it can be a delicate therapeutic task to remove such inhibitions in a person with a great deal of repressed hostility without precipitating either a psychotic break or excessive acting out.

Postoffense treatment on the other hand must cope not only with the problem of guilt, but also with limitations imposed by judicial procedures. If an extremely assaultive offense has been committed, it is likely that the patient will have to be treated in some form of penal institution. As noted above, the program of such institutions is to reward control and conformity and to punish assertiveness or aggression. This means that the goals of the institutional program and the therapeutic program will be at complete odds with each other. The patient will have few chances to practice assertive and mildly aggressive responses in a setting in which they are apt to be rewarded.

If an attempt were made to match the treatment program to the needs of the different types of inmates within a given institution, chaos would result. Undercontrolled Aggressive people would be punished for doing the same sorts of things that Chronically Overcontrolled people were being encouraged to do. This would naturally be interpreted as injustice and favoritism. It would, therefore, be necessary to treat the two types of offenders separately, either at different institutions or by incarcerating the Undercontrolled offender while placing the Chronically Overcontrolled person on probation with outpatient therapy. However, since the Chronically Overcontrolled assaultive person is likely to have committed the more severe offense, it would be very difficult to obtain support either from the public or from legislative bodies for such a program.

The proposed typology has implications for psychological theory as well as for clinical practice. The model of aggressive dynamics which has been used throughout this investigation is the frustration-aggression model originally proposed by Dollard et al. (1939), in which instigation to ag-

gression was viewed as the result of frustration. Whether or not this instigation resulted in overt aggressive behavior depended upon the relative balance of instigation and inhibition. If the inhibitory forces exceeded the instigation to aggression, no aggressive response should result; on the other hand, if instigation exceeded inhibition, in the presence of sufficient cues to aggression, then overt aggressive behavior was likely.

Attempts have been made to apply this schema to the prediction of the relative intensity of the overt aggressive response. The most frequent hypothesis is that the intensity of the aggressive response is a function of the net strength of instigation minus inhibition. (If inhibition exceeds instigation, the result is less than zero and no response occurs.) Bandura and Walters (1959) write for instance:

> By subtracting the height of the curve representing the strength of the inhibitory response from the height of the curve representing the strength of the inhibited response, it is possible to represent the strength of the overt response that may be expected at any point on the dissimilarity continuum [p. 133].

This formulation, however, does not seem applicable to extremely assaultive, Chronically Overcontrolled offenders for whom the violence of the overt act is out of all proportion to the immediate stimulus. In one case for instance, a mild-mannered inoffensive bachelor was abused and insulted for weeks by an Undercontrolled Aggressive neighbor who had a long record of assaultive behavior. One night, after enduring 45 minutes of abuse, he found one final insult too much to bear and proceeded to shoot his tormentor four times at close range. Immediately prior to that final insult, his instigation to aggression was apparently less than his inhibitions, for he made no aggressive response. The amount of instigation added by the next taunt was sufficient to increase the level of instigation to the point where it finally exceeded the inhibitions and the violent aggressive response was elicited.

In this case, then, the *net* strength of instigation minus inhibition was undoubtedly quite low and the assault which took place was out of all proportion to it. On the other hand, the *absolute* level of instigation to aggression, which had been slowly building up over the months, was probably quite high. In this case, then, it would appear that the strength of the overt aggressive act was a function of the total amount of instigation and not the net strength of instigation minus inhibition. In fact, if the intensity of the aggressive act were a function of the net strength, then one would predict that no person with excessive inhibitions could ever commit more than the mildest of aggressive acts, except under conditions of extreme provocation. However, the data in the present study indicate that the opposite is true; when excessively controlled people do aggress it is more likely to be in an extreme fashion with inadequate provocation. (See Appendix B.)

This notion that the violence of the overt aggressive act is a function of the total instigation to aggression, rather than the net strength, has, of course, been implicit during our whole discussion of the Chronically Overcontrolled offender, for without it his overwhelming violence with minimal external provocation is incomprehensible. However, while this formulation neatly accounts for the paradoxically more extreme aggression of the Chronically Overcontrolled type as opposed to the Undercontrolled Aggressive type, it is oversimplified. In essence it reduces to: *The greater the instigation to aggression toward a target, the greater the degree of violence of the aggressive response to the target, if an aggressive response is allowed to occur.* The difficulty with this formulation is that it overlooks the phenomenon of response generalization. A person who is strongly angered by his wife may be motivated to make a highly aggressive response, but, because of excessive inhibitions, suppress the response. One way he could deal with the situation would be to displace his aggression to another target (Miller, 1948). Another way would be to make a lesser aggressive response to the original target. Thus while his original inclination may have been to hit his wife, he may suppress this response and instead make a sarcastic

remark, slam the door, or behave with excessive politeness in a passive-aggressive manner. Thus, for any given target there is a constant level of instigation to aggression but any number of possible aggressive responses. Obviously, then, there is no simple direct relationship between instigation, inhibition, and overt aggressiveness.

This study has therefore highlighted two areas of oversimplification. The first was the oversimplified notion that aggressive behavior is associated only with deficient controls. The second was that relative balance of inhibition and aggression was sufficient to account for the strength of the aggressive response in all situations. There is also a third area of thought which seems oversimplified in the light of the present investigation. This is the area of theories about the etiology of delinquent behavior. The public typically asks psychologists, sociologists, and educators, "What causes delinquency?" and, all too often, they have been willing to respond with some single explanation. Sometimes this explanation is complex such as "anomie" (Merton, 1957) or "superego lacunae" (Johnson, 1949). At other times it may be quite simple such as "too much violence on television."

If nothing else, this study should demonstrate that any attempt to establish a single, simple cause for crime or delinquency is certain to fail. It is apparent that even within the relatively simple category of aggressive behavior there are vast differences in personality patterns among the people who engage in such behavior. If we expand the horizon to include the whole panorama of illegal behavior subsumed under the headings of, "crime" or "delinquency," ranging from dope peddling to income tax evasion, from safecracking to homosexuality, from traffic violations to murders, the futility of finding a single cause or a single cure can be seen. The first step needed is adequate classification based on empirical research; the next is study of the dynamics of each type or class to determine the appropriate treatment; the final step is applying the research so that instead of making the punishment fit the crime we can instead make the treatment fit the criminal. The present study represents one beginning to this first task of empirical classification.

APPENDIX A

Ten-Point Scale of Aggressiveness on Which Assaultive Offenders Were Rated

Scale Value	Behavior
1	Subject showed good restraint. Resorted to aggression only when it was clearly dictated by circumstances, that is, hit back with equal or less force; self defense.
2	Less restraint shown but degree of aggression still quite appropriate; or instrumental aggression (i.e., aggression whose primary motive is something other than inflicting pain—strong-arm robbery), with enough violence to accomplish the end goal, but no more.
3	Aggression exceeds provocation, but not inappropriate in subculture; or instrumental aggressive acts where degree of violence begins to indicate that desire to inflict pain is also a motive.
4	Aggression exceeds provocation even more but would not be viewed as a particularly extraordinary response by members of subculture—hitting person who calls defendant a name or ganging up on victim; or instrumental aggression which clearly exceeds amount needed to accomplish act.
5	Acts of aggression clearly motivated by desire to inflict pain or injury. Culture and situation less supportive of degree of violence used. Would probably be rejected by adult members of subculture but not necessarily by peer group, for example, hitting when down. Violence at this point still not likely to seriously or permanently injure victim, although severe injuries might occur accidentally.
6	Even less justification than (5)—victim weaker or frailer. More apt to do serious harm (stomping), or use of weapon versus superior, unarmed antagonist.
7	Serious aggression with inadequate provocation. Apt to result in serious injury to

victim. Most members of subculture would feel use of this much violence in this situation unjustified, although it might still be sufficiently provocative to call for a lesser physical response such as use of weapon when called name or in gang fight versus unarmed opponents of equal or less size.

8 More serious aggression. Death, or permanent disability quite likely. There may be some external motivation apparent for act, but it clearly does not justify this degree of response.

9 Extremely severe aggression with serious consequence probable. Would be rejected by all in subculture as unjustified. Some glimmer of external motivation still apparent, for example, a murder or assault with a deadly weapon with little motivation, but in heat of anger.

10 Completely externally unprovoked, extremely serious aggression with extreme physical harm probable. No external motivation, for example, a "senseless" murder or assault with a deadly weapon, not even done in the heat of anger.

APPENDIX B

Summary of Offenses Committed by the Assaultive Subjects

Extremely Assaultive Offenses

Rating: 9.0–10.0

S 07009 Shot and killed mother with rifle from ambush. Also fired at father but missed. No known external provocation.

S 02391 Shot a strange woman with a rifle while cruising in a car with two friends looking for youths who had allegedly beaten him up.

S 04116 Fired at but missed an adult who had threatened to slap his face the day before. Later stated that he intended to kill rather than frighten the victim.

Rating: 8.0–8.9

S 49797 Tried to talk 19-year old housewife into letting him enter her home. When told to leave he grabbed her arm, pulled her through the chain-locked door, and hit her on the head with a gun. Upon gaining entrance he fought with her and then ran away.

Rating: 7.0–7.9

S 98083 Defendant hit the victim in the eye twice, knocked him down and continued to hit the victim with both fists until blood stained his pants. Victim maintained there was no provocation; the defendant said the victim had called his mother names and had dared him to hit him.

S 72044 Defendant harassed the victim verbally. The victim hit the defendant whereupon the defendant went home, secured a knife, returned and slashed the victim.

S 92552 Defendant felt (with some possible justification) that his teacher was persecuting him. He responded with passive aggression, but when he discovered he had been suspended, he returned and clubbed the teacher on the head with a mummified deer hoof.

Rating: 6.0–6.9

S 45091 In the course of a violent family fight in which the defendant's brutal and intoxicated father was beating his mother, the defendant secured the father's pistol. When the father turned on him, his sister shouted at him to fire. He did so, killing the father.

S 97653 (See also 94887) Three drunken white boys had harassed a group of small Negro children. The defendant was one of a group of large Negro boys who then started a fight with the white boys. When one of the white boys threatened the defendant with a tennis racket, he pulled a zip gun and fired it. No one was hit.

MODERATELY ASSAULTIVE OFFENSES

RATING: 5.0–5.9

S 44125 The defendant was one of three Negro boys who tried to taunt three white boys into a fight. When the white boys did nothing, the defendant hit one of them in the jaw with brass knuckles, chased and kicked him.

S 07232 The victim, during a school class, told the defendant he was an "ass." Outside the classroom the defendant grabbed the victim and hit him in the mouth with brass knuckles.

S 88663 Two counts of battery. No. 1: When his brother was in a fight the defendant assisted by attacking the opponent with a wrench. No. 2: Was the principal aggressor when he and his friends attacked some other young boys who were smaller than they were.

S 03805 Argued with large football player who offered to fight him. Armed self with a broken bottle and waited for the football player. The victim got a long stick, and they attacked each other.

S 92441 After the victim and defendant had boxed in school, the rumor spread that the victim had challenged the defendant to an after-school fight. Later in a group, the victim denied this challenge. The defendant suddenly hit the victim and knocked him down. The defendant's friends joined in and someone kneed and kicked the victim in the face.

RATING: 4.0–4.9

S 45986 Fighting broke out among a large group of teenagers. The victim tried to leave the group, and the defendant chased him and hit him several times as he tried to get away.

*S*s 18067 and 82579 A 51-year old man observed a group of juveniles on his fence. He told the group to disperse, turned his back and the *S*s attacked him knocking him to the ground causing 15 stitches to be required. Both these defendants were involved.

S 67614 The defendant walked up to victim, called him a "fink," and knocked him out, with his fist. He possibly also kicked the victim when down. His excuse was that the victim had once reported the defendant's brother for some offense.

S 42363 The defendant and friends demanded a sailor give them a dime. The sailor ignored them, so the defendant pushed the sailor, swung at him, and started a fight.

S 32410 In the course of an attempt to snatch a woman's purse, the defendant struck the woman on the head with a blunt object without provocation.

S 06750 The defendant was one of a group of boys aged 8 to 11 who stole the wallet from an 82-year-old cripple and then taunted and jostled him and also pushed an adult female who attempted to assist.

S 53254 After the victim's older brother had beaten up a friend of the defendant, the defendant and his friends harassed the victim's entire family. In this instance, he punched the victim without provocation. The defendant maintained the victim had called him a "black nigger."

S 53279 After crashing a party, the defendant got into an argument with a larger boy. He pulled out a can-opener, planning to threaten him, but in the ensuing scuffle, the victim was scratched.

S 96458 The defendant was one of 70 teenagers who milled around and blocked traffic. The defendant hit a 52-year-old man in one of the cars in the face. When a peacemaker tried to interfere, he and his friends chased him home, tore the clothes off him and wrecked his house.

RATING: 3.0–3.9

S 35148 The defendant struck the victim who was walking along the street. He claimed that the victim had bumped into him, but the victim denied this.

S 60952 The defendant asked the victim for a nickle. The victim refused so the defendant shoved him against some lockers. Three stitches had to be taken in the victim.

S 06347 In the course of brandishing a small pocket knife, the defendant inflicted a small stab wound in a boy's back under possibly accidental circumstances.

RATING: 2.0–2.9

S 22841 In the course of a purse snatching in which the victim was knocked down, this defendant kicked the victim in the face but inflicted no damage. He was wearing sneakers and said that his foot slipped.

S 94887 (See also 97653) This defendant was another member of the group of Negroes who were in a fight with white boys who had been harassing some smaller children. This particular defendant was only minimally involved in the altercation.

S 69900 The defendant and a friend accosted the victim in the schoolyard. The defendant and the victim had been feuding for some time. When the victim made a gesture as if he was going to hit the defendant, the defendant struck first, hard, causing a concussion.

APPENDIX C

BEHAVIOR CHECK LIST[C1] ON WHICH COUNSELORS REPORTED INSTANCES OF AGGRESSIVE BEHAVIOR ON THE THIRD AND TENTH DAYS OF DETENTION

Instructions: In making these reports check those types of aggressive behavior you have noticed *during that particular period.* Be sure to look for the less obvious and more secretive or subtle types of aggressive behavior, as well as the more obvious kinds. It is important that you turn in these sheets at the end of the third and tenth day as they are completed.

Check List:

1. *Physical Attack.* Starting fights, hitting, pushing—unprovoked by verbal and physical attack of other children.
2. *Bragging.* Assertively, with show of bravado—"I can do this better than you" sort of thing.
3. *Threatening.* Specific, hostile verbal or physical threat, or threatening act.
4. *Teasing.* Including specific acts which appear designed to annoy or irritate, hurt, or humiliate.
5. *Saucy, Impertinent.* "Smart-alecky."
6. *Insulting, Name Calling.* Direct face-to-face with object of hostility.
7. *Ridiculing, Mocking, Making-Fun-Of.*
8. *Bullying.* Another who is smaller or weaker, or who for some reason can't defend himself effectively.
9. *Verbal Castigation.* Cursing, upbraiding, blaming, "giving somebody hell" verbally.
10. *Malicious Gossip, Depreciating, Defaming,* or *Tale Carrying (Tattle-Taling).*
11. *Destructive.* Breaks things, defaces walls, tears or dirties clothing or bedding, etc.
12. *Temper Tantrums.* Fits of rage, screams, kicks, scratches, etc.
13. *Running Away.*
14. I have not observed this boy at all due to my schedule or due to the fact he has been in isolation the whole time.[C2]

[C1] Check list originated by Naylor (1952) and reproduced with his permission.

[C2] This item was added by the present investigator and did not appear on the original Naylor check list.

APPENDIX D

RATING SCALE[D1] FILLED OUT BY COUNSELORS ON THIRD AND TENTH DAYS OF DETENTION

Instructions: Check the point on each scale which in your opinion best describes the behavior of this child during the past week.

In making these ratings, try to compare him with all the other children you have known. Judge him with respect to each quality *independently;* that is, judge objectively and try not to be influenced by your general impression of him.

[D1] Rating scale originated by Naylor (1952) and reproduced with his permission.

You may check *any* place on a scale.
Be sure to rate him on all five scales.
Each child is to be given a rating on the third and tenth day of custody.

Scale 1. Uncooperative-Cooperative

1.	2.	3.	4.	5.
Extremely uncooperative; refuses to follow any suggestions; unwilling, antagonistic.	Uncooperative: replies perfunctorily to questions; indifferent.	Takes situations for granted; responds willingly but volunteers little.	Likes being asked to do things; volunteers occasionally.	Very cooperative. Volunteers help readily; anxious to do anything asked.

Scale 2. Amiable-Quarrelsome

1.	2.	3.	4.	5.
Actively dislikes quarrels. Acts as peacemaker. Good humored.	Has sunny disposition. Quarrels less than average.	Quarrels under real provocation; occasionally starts quarrel. Generally amiable.	Quarrels more than average child.	Pronounced tendency to be quarrelsome; has a "chip on the shoulder."

Scale 3. Aggressive-Submissive

1.	2.	3.	4.	5.
Threatens others; dominant; reacts to reproof violently; overtly aggressive; starts trouble.	Seldom or reluctantly gives in; reacts to violence with violence. Threatens others.	Complies with normal authority; reacts with violence only when provoked.	Gives in readily; objects to violence with "Stop!" but not with blows.	Complies with all requests; submits to violence without doing anything about it.

Scale 4. Docile-Rebellious

1.	2.	3.	4.	5.
Passively agrees to everything; no sign of resistance or unwillingness.	Tends to accept suggestions and do what he is told without resistance.	Conforms normally to all reasonable requests and accepts authority as necessary.	Tends to resist authority but will conform if enough pressure is put on him.	Hostilely defiant rejects all suggestions and resists any restraint.

Scale 5. Antagonistic-Friendly

1.	2.	3.	4.	5.
Marked hostility, suspiciousness, or unfriendliness.	Not as marked as 1, but less friendly than the average child.	About like the average. Has both likes and dislikes.	More friendly and outgoing than the average child, but not as marked as 5.	Exceptionally outgoing and friendly. Likes practically everyone and wants them to like him.

APPENDIX E

TABLE E1

LISTS OF "OVERCONTROLLED" AND "UNDERCONTROLLED" ADJECTIVES FROM THE GOUGH ADJECTIVE CHECK LIST

Overcontrolled		Undercontrolled	
Adjective No.	Adjective	Adjective No.	Adjective
28	Cautious	7	Aggressive
43	Conscientious	14	Argumentative
45	Considerate	17	Assertive
49	Cooperative	23	Boastful
85	Fearful	24	Bossy
100	Gentle	52	Cruel
111	Helpful	59	Demanding
129	Inhibited	70	Dominant
146	Mannerly	114	Hostile
149	Meek	121	Impulsive
158	Nervous	138	Irritable
171	Peaceable	144	Loud
191	Quiet	152	Mischievous
207	Retiring	168	Outspoken
214	Self-controlled	188	Quarrelsome
230	Shy	197	Rebellious
253	Submissive	210	Rude
268	Timid	211	Sarcastic
297	Withdrawn	228	Show-off
299	Worrying	271	Tough

REFERENCES

BANDURA, A., & WALTERS, R. H. *Adolescent aggression—the influence of childtraining practices and family interrelationships.* New York: Ronald Press, 1959.

BERG, I. A., & FOX, V. Factors in homicides committed by two hundred males. *Journal of Social Psychology,* 1947, **26,** 109–119.

BERKOWITZ, L. *Aggression: A social psychological analysis.* New York: McGraw-Hill, 1962.

DEMING, R. W. Reactions to frustration of assaultive delinquents. Research report No. 2, Alameda County Probation Department, December, 1960.

DOLLARD, J., DOOB, L. W., MILLER, N. W., MOWRER, O. H., & SEARS, R. R. *Frustration and aggression.* New Haven: Yale Univer. Press, 1939.

ELIZUR, A. Content analysis of the Rorschach with regard to anxiety and hostility. *Journal of Projective Techniques,* 1949, **13,** 247–284.

FINNEY, B. C. Rorschach test correlates of assaultive behavior. *Journal of Projective Techniques,* 1954, **18,** 247–256.

GORLOW, L., ZIMET, C. N., & FINE, H. J. The validity of anxiety and hostility Rorschach content scores among adolescents. *Journal of Consulting Psychology,* 1952, **16,** 73–75.

GOUGH, H. G. *Manual for the California Psychological Inventory.* Palo Alto, Calif.: Consulting Psychologists Press, 1960.

GUILFORD, J. P. *Fundamental statistics in psychology and education.* (3rd ed.) New York: McGraw-Hill, 1956.

HOLTZMAN, W. H. *Holtzman inkblot technique administration and scoring guide.* New York: Psychological Corporation, 1961.

JOHNSON, ADELAIDE. Sanction for superego lacunae of adolescents. In K. R. Eissler (Ed.), *Searchlights on delinquency.* New York: International Univer. Press, 1949.

KAHN, M. W. A comparison of personality, intelligence, and social history of two criminal groups. *Journal of Social Psychology,* 1959, **49,** 33–40.

KLOPFER, B., AINSWORTH, MARY D., KLOPFER, W. G., & HOLT, R. R. *Developments in the Rorschach technique.* Vol. 1. Yonkers, N.Y.: World Book, 1954.

KOSTLAN, A. A method for the empirical study of psychodiagnosis. *Journal of Consulting Psychology,* 1954, **18,** 83–88. Reprinted in E. I. Megargee (Ed.), *Research in clinical assessment.* New York: Harper & Row, in press.

LAMBERTI, J. W., BLACKMAN, N., & WEISS, J. M. A. The sudden murderer: A Preliminary report. *Journal of Social Therapy,* 1958, **4,** 2–15.

LINDZEY, G., & TEJESSEY, CHARLOTTE. Thematic Apperception Test: Indices of aggression in relation to measures of overt and covert behavior. *American Journal of Orthopsychiatry,* 1956, **26,** 567–576.

LITTLE, K., & SHNEIDMAN, E. Congruencies among interpretations of psychological tests and anam-

nestic data. *Psychological Monographs,* 1959, **73,** No. 6 (Whole No. 476). Reprinted in E. I. Megargee (Ed.), *Research in clinical assessment.* New York: Harper& Row, in press.

MEGARGEE, E. I. The utility of the Rosenzweig Picture-Frustration Study in detecting assaultiveness among juvenile delinquents. Paper read at Southwestern Psychological Association, San Antonio, Texas, April 1964.

MEGARGEE, E. I. The relation between inner controls and the overt expression of aggression. Paper read at Southwestern Psychological Association Convention, Oklahoma City, Oklahoma, April 1965.

MEGARGEE, E. I., & MENDELSOHN, G. A. A cross validation of twelve MMPI indicies of hostility and control. *Journal of Abnormal and Social Psychology,* 1962, **65,** 431–438. Reprinted in E. I. Megargee (Ed.), *Research in clinical assessment.* New York: Harper & Row, in press.

MEGARGEE, E. I., & MENDELSOHN, G. A. The assessment and dynamics of murderous aggression: A progress report. Paper read at California State Psychological Association, San Francisco, December 1963.

MERTON, R. *Social theory and social structure.* (Rev. ed). Glencoe, Ill.: Free Press, 1957.

MILLER, N. E. Theory and experiment relating psychoanalytic displacement to stimulus-response generalization. *Journal of Abnormal and Social Psychology,* 1948, **43,** 155–178.

MURSTEIN, B. I. The projection of hostility on the Rorschach and as a result on ego threat. *Journal of Projective Techniques,* 1956, **20,** 418–428.

MUSSEN, P. H., & NAYLOR, H. K. The relationship between overt and fantasy aggression. *Journal of Abnormal and Social Psychology,* 1954, **49,** 235–240. Reprinted in E. I. Megargee (Ed.), *Research in clinical assessment.* New York: Harper & Row, in press.

NAYLOR, H. K. The relation between overt aggression and aggression and punishment expressed in the Thematic Apperception Test. Unpublished Masters thesis, Ohio State University, 1952.

Newsweek, June 21, 1965, p. 29 ff.

PATTIE, F. A. The effect of hypnotically induced hostility on Rorschach responses. *Journal of Clinical Psychology,* 1954, **10,** 161–164.

PURCELL, K. The TAT and anti-social behavior. *Journal of Consulting Psychology,* 1956, **20,** 449–456.

RADER, G. E. The prediction of overt aggressive verbal behavior from Rorschach content. *Journal of Projective Techniques,* 1957, **21,** 294–306.

SCHULTZ, L. G. The wife assaulter: one type observed and treated in a probation agency. *Journal of Social Therapy,* 1960, **6,** 103–111.

SIEGEL, S. *Nonparametric statistics for the behavioral sciences.* New York: McGraw-Hill, 1956.

SOMMER, R., & SOMMER, DOROTHY T. Assaultiveness and two types of Rorschach color responses. *Journal of Consulting Psychology,* 1958, **22,** 57–62. Reprinted in E. I. Megargee (Ed.), *Research in clinical assessment.* New York: Harper & Row, in press.

STEARNS, A. W. Murder by adolescents with obscure motivation. *American Journal of Psychiatry,* 1957, **114,** 303–305.

STONE, H. The relationship of hostile-aggressive behavior to aggressive content on the Rorschach and Thematic Apperception Test. Unpublished doctoral dissertation, University of California, Los Angeles, 1953.

STORMENT, CHARLYNE T., & FINNEY, B. C. Projection and behavior: a Rorschach study of assaultive mental hospital patients. *Journal of Projective Techniques,* 1953, **17,** 349–360.

TOWBIN, A. Hostility in Rorschach content and overt aggressive behavior. *Journal of Abnormal and Social Psychology,* 1959, **58,** 312–316.

WALKER, R. G. A comparison of clinical manifestations of hostility with Rorschach and MAPS tests performances. *Journal of Projective Techniques,* 1951, **15,** 444–460.

WECHSLER, D. *Manual for the Wechsler Intelligence Scale for Children.* New York: Psychological Corporation, 1949.

WECHSLER, D. *Manual for the Wechsler Adult Intelligence Scale.* New York: Psychological Corporation, 1955.

WEINBERG, W. L. The relationship of the extrapunitive category of the picture-frustration study to an independent criterion of aggression in prisoners. Unpublished master's thesis, University of Oregon, 1953.

WEISS, J. W. A., LAMBERTI, J. W., & BLACKMAN, N. The sudden murderer. *Archives of General Psychiatry,* 1960, **2,** 669–678.

WICKHAM, W. Psychological aspects of teenage murderers. Report to the Juvenile Officers Coordinating Council of Alameda County, San Leandro, California, November 20, 1956.

WOLFGANG, M. E. Victim-precipitated criminal homicide. *Journal of Criminal Law, Criminology and Police Science,* 1957, **48,** 1–11.

YOUNG, FLORENE M. Responses of juvenile delinquents to the Thematic Apperception Test. *Journal of Genetic Psychology,* 1956, **88,** 251–259.

(Received May 3, 1965)

Vol. 80, No. 4 Whole No. 612

Psychological Monographs: General and Applied

GENERALITY OF WORD-ASSOCIATION RESPONSE SETS[1]

LOUIS J. MORAN

University of Texas

The tendency of some Ss to give predominately 1 category of associate in the word-association experiment, regardless of word list, was examined in a series of studies. Earlier findings on such tendencies were confirmed with several samples of college students and with Spanish-speaking Ss. A 4th such tendency, or idiodynamic set, perceptual-referent (Jung's "predication type"), was added to the 3 sets previously described. Evidence was presented for the sets as more generally representing bases for matching word pairs, forming a hierarchy of increasing linguistic sophistication in the order: perceptual referent, object referent, concept referent, and dimension referent. Word-association commonality was discussed as an arbitrary average across several stable subhierarchies.

IN the "free" word-association experiment, where the subject is instructed to give the first word that occurs to him when he hears the stimulus word, some individuals tend to give predominately one "category" of associate. Because these individuals are highly consistent in giving one characteristic category of associate on different occasions and to different word lists, they are said to manifest enduring "idiodynamic" associative sets. Three such sets recently have been described (Moran, Mefferd, & Kimble, 1964). Individuals with an object-referent set tend to give associates categorized as "functional," usually naming another object that is associated in everyday experience with the object denoted by the stimulus word, for example, FOOT, shoe; BOAT, dock. Other individuals evidence a concept-referent set, characterized by a preponderence of synonym (e.g., SMALL, little) and superordinate (e.g., CABBAGE, vegetable) associates. A third group of individuals tends to give very fast contrast (e.g., BLACK, white) and logical coordinate (e.g., APPLE, orange) associates and are said to have a set for speed in responding.

Dependent variables in the free word-association experiment are significantly influenced by an interaction between the subject's idiodynamic set and the set compatibility of the stimulus words. Response faults (e.g., delayed reaction time, blank, multiword, etc.) occur less frequently to stimulus words that are most compatible with the subject's set. Commonality score (degree to which a subject's associates correspond to those of a normative group) is a partial function of the number of stimulus words in the list that happen to be compatible with the subject's set. Grammatical form of associates is influenced by set: object-referent, concept-referent, and speed sets tend to produce noun, verb, and adjective associates, respectively (Moran et al., 1964).

The studies which follow were designed to determine the generality and reliability of specific idiodynamic associative sets and to seek a unifying rationale for these individual differences in linguistic habits.

SET-COMPATIBLE WORD LISTS

The pervasive influence of associates sets needs to be taken into account in the construction of word lists. To illustrate, of the 100 primary (most popular) response words in the Kent-Rosanoff list, 38 are contrast or coordinate associates, yielding a sum commonality score of 1,746. Only 10 primary response words in the list are synonym or superordinate associates, yielding a sum commonality score of 351 (Russell & Jen-

[1] This investigation was supported by Research Grant MH-08778 from the National Institutes of Health, United States Public Health Service.

kins, 1954). Because of this inequality, it may be predicted that subjects with a concept-referent set will commit more faults and achieve a lower commonality score on the Kent-Rosanoff list than will subjects with a so-called "speed" set. Indeed, even the seemingly well-established speed set of subjects who give many contrast-coordinate associates may be an artifact of biased word lists. Studies that have reported a very marked increase in contrast associates induced by time pressure (e.g., Siipola, Walker, & Kolb, 1955) have used word lists that were "loaded" with popular contrast-evoking stimulus words. Time pressure may well increase the frequency of popular associates (Horton, Marlowe, & Crowne, 1963); but whether these popular associates turn out to be predominately contrasts, or synonyms, or some other category is a function of the word list selected by the experimenter. The same ambiguity attends some interpretations of alleged changes in category of associations in developmental studies. As children approach adulthood they will, by definition, tend to give more adultlike contrast associates to the Kent-Rosanoff list. It should be noted, however, that to a different word list, one that evoked predominately synonym associates from adults, the same maturing children would tend, by definition, to give more adultlike synonym (rather than contrast) associates.

By a careful selection of stimulus words, lists may be constructed that are equally compatible with two or more associative sets. For example, words like BLOSSOM, which evoke predominately synonyms (flower, 672) may be counterbalanced in the same list with words like GIRL, which evoke contrast associates with nearly the same frequency (boy, 670) (Russell & Jenkins, 1954). Such a list should be reasonably compatible with either a set to give synonyms or a set to give contrasts. The frequency tables for 400 stimulus words, based upon the associations of 196 men (Moran et al., 1964), were used in this manner to construct two word lists, equally compatible with the three sets described above.

It was possible to locate among the 400 words, 20 words which had elicited associates with near equal frequency in all three categories: for example, HEAT elicited fire, warm, and cold with nearly the same frequency. These words were assigned 10 to List A and 10 to List B, so that overall set compatibility was maintained within each list and between lists.

An additional 24 stimulus words were found which had elicited, with nearly equal frequency, responses compatible with two of the sets, but not the third; for example, BEAUTIFUL elicited pretty and ugly, but no associates categorized as object referent. These words were assigned 12 to List A and 12 to List B, so that all three sets were equally represented within and between lists.

Of the remaining 356 stimulus words in the pool, 36 had elicited predominately a response word compatible with one set only: for example, DARK elicited mainly light. Lists A and B each received 18 of these words, matched overall for compatibility with the three sets. The last 18 words in each list were ordered systematically, with respect to the predominate category evoked, that is, contrast-coordinate, synonym-superordinate, functional, contrast-coordinate, etc.

The two 40-word lists, then, each contained Words 1–10 equally compatible with all three sets, Words 11–22 each compatible with two sets but not the third, and Words 23–40 each compatible with only one of the three sets. The two lists were carefully matched for overall compatibility with all three sets. The complete word list is provided in the Appendix.

Subjects

Two large classes in freshman psychology at the University of Texas, totaling 482, served as subjects.

Administration of Word Lists

Subjects were instructed to write the first word that came to mind when they heard the word read by the examiner over the auditorium speaker system. They were told that the words would be read at 5-second intervals and to leave a blank space if no response word came to mind. Also, they were asked not to change a response or to return

to fill in a blank space later. Words were read in the order: List A, Words 1–10; B, 1–10; A, 11–22; B, 11–22; A, 23–40; B, 23–40. The same examiner tested all subjects.

Variables

The following variables were scored. Most of the scoring was clerical, using a manual consisting of the prescored responses of 196 men (Moran et al., 1964).

1. *Functional.* Stimulus word and response word each separately denote entities or processes between which there is an explicit functional relationship, for example, FOOT, shoe.

2. *Synonym or superordinate.* Synonym—response word has exactly the same meaning as the stimulus word in one or more ordinary and appropriate contexts, for example, BLOSSOM, flower. Superordinate—stimulus word denotes an immediate member of the class or category denoted by response word, for example, CABBAGE, vegetable.

3. *Contrast or logical coordinate.* Contrast—response word negates or contrasts with the meaning of stimulus word in one or more ordinary and appropriate contexts: for example, DARK, light. Logical coordinate—stimulus word and response word separately denote immediate members (of equal logical order) of the same class or category, for example, BLUE, yellow.

4. *Total faults.* Sum of blanks and multiwords.

5. *Commonality.* Sum frequency of subject's associates also given by the 196 men: that is, each response word received a "score" corresponding to the number in the normative group that gave the same response word.

Statistical Analyses[2]

All correlations in this report are Pearson product-moment correlation coefficients. All factor analyses are for Principal Components, rotated by the normalized varimax method, with unities placed in the diagonal. Factor extraction in all analyses was stopped when eigenvalues dropped below unity.

Equated Word Lists

Means, standard deviations, and intercorrelations for Lists A and B, based upon the 482 college students, are given in Table 1. Here it may be seen that the mean frequency of functional and of contrast-coordinate associates was nearly equal in both lists, with synonym-superordinate associates somewhat less frequent on both lists. Internal consistencies of the variables ranged from .56 to .84, after correction for an 80-word list. List A was less "difficult" (higher commonality and fewer faults) than List B.

The five variables for List A and for List B were factor analyzed jointly, as shown in Table 2. Factor I represents the contrast-coordinate associates in both lists, loading .75 in List A and .74 in List B. Functional associates are represented by Factor II, loading .76 in List A and .83 in List B. Factor III represents the synonym-superordinate associates, which loaded .84 and .86 in Lists A and B, respectively. Thus, the same three factors found in earlier samples (Moran et al., 1964) appeared also in the present college sample.

The consistent appearance of these three factors in samples of subjects tested on 4 successive days, with different word lists each day, led to the initial postulation of the three idiodynamic sets described above. In a table much like Table 3, below, it was demonstrated that the subjects who had evidenced a definite set had achieved a higher commonality score and had committed the fewest faults on the stimulus words most compatible with their set (Moran et al., 1964). For the present sample of 482 students, predictions were made from associates on one group of stimulus words to an independent group of criterion stimulus words. Subjects were selected to represent a set if they had a standard score greater than 1.00 on one of the three set variables, and less than .00 on the other two set variables, based upon their first 44

---

[2] Computations were carried out at the Computation Center of the University of Texas, with programs compiled by Don Veldman, of the Department of Educational Psychology, to whom appreciation is expressed for his very generous and helpful consultations on statistical problems.

TABLE 1

Means, Standard Deviations, and Intercorrelations of 10 Variables for 482 Freshmen

Variable	Mean	SD	Intercorrelations								
			1	2	3	4	5	6	7	8	9
List A											
1. Functional	9.8	3.0									
2. Synonym-superordinate	7.4	2.8	−.08								
3. Contrast-coordinate	10.0	4.2	−.33	−.07							
4. Total faults	3.0	2.5	−.17	−.28	−.33						
5. Commonality	627.2	151.5	.14	.34	.55	−.44					
List B											
6. Functional	9.0	3.3	.39	−.06	−.24	−.13	−.02				
7. Synonym-superordinate	7.8	3.1	.02	.53	−.05	−.10	.23	−.24			
8. Contrast-coordinate	9.3	4.5	−.22	−.02	.73	−.21	.49	−.34	−.08		
9. Total faults	3.5	2.8	−.05	−.16	−.24	.53	−.30	−.16	−.24	−.29	
10. Commonality	585.2	151.1	.07	.15	.46	−.30	.57	.17	.25	.57	−.40

associates only (Words 1–22 in Lists A and B). It will be recalled that the last 18 words in each list consisted of three groups of 6 words each which had elicited (from 196 men) predominately associates compatible with only one set. The last 18 words in both lists were combined so that each type of stimulus word was represented by 12 set-specific stimulus words. It was predicted that subjects who evidenced a set on the first 44 words would achieve their highest commonality score on the group of criterion stimulus words most compatible with their set. Faults were too infrequent (mean below one) on these high-commonality words to permit analysis.

Note first in Table 3 that *average* commonality values obtained on the three groups of stimulus words remained equated for the present sample, being 26.0, 25.5, and 25.9. As predicted, however, subjects achieved their highest commonality scores on stimulus words most compatible with their own set. The commonality score of an individual thus is a partial function of the interaction between his idiodynamic associative set and the set compatibility of the stimulus words. To illustrate, the mean com-

TABLE 2

Normalized Varimax Rotated Factors for 482 Freshmen

Variable	Factor I	Factor II	Factor III	h^2
List A				
1. Functional	.05	.76	.01	.58
2. Synonym-superordinate	.15	−.02	.84	.73
3. Contrast-coordinate	.75	−.47	−.23	.83
4. Total faults	−.64	−.29	−.22	.54
5. Commonality	.78	−.01	.24	.66
List B				
6. Functional	.07	.83	−.19	.72
7. Synonym-superordinate	.10	−.09	.86	.76
8. Contrast-coordinate	.74	−.48	−.22	.83
9. Total faults	−.62	−.23	−.21	.48
10. Commonality	.79	.06	.09	.63

TABLE 3
EFFECT OF SET UPON COMMONALITY SCORES

N	Type of subject	Type of stimulus word: Functional	Synonym-Superordinate	Contrast Coordinate	Mean
		(Average commonality [%] of responses)			
41	(F) Object-referent	28.1	21.5	23.5	24.4
25	(SS) Concept-referent	25.4	28.6	20.6	24.9
39	(CC) Speed set	24.4	26.5	33.7	28.2
	Mean	26.0	25.5	25.9	

Note.—Type of subject represented by subjects with z score greater than 1.00 on variables representing one set and less than .00 on variables representing the other two sets, based upon first 44 associates. Type of stimulus word each represented by 12 words that had elicited predominately one type of response from 196 men; these were the last 36 stimulus words in the list.

monality score (i.e., *averaged* across the three types of stimulus words) of object-referent and concept-referent set subjects was almost the same, 24.4 and 24.9, respectively. But if all stimulus words had been of the "functional" type, the object-referent set subjects would have achieved a 13% higher commonality score; if all had been of the synonym-superordinate type, a 12% lower score. Thus, the type of stimulus words in a list could make a 25% difference in the commonality score achieved by these subjects.

The present word lists proved to be reasonably well equated for set compatibility, though split-half reliability and fault-evoking potential was fairly low. Since faults (as well as reaction time, i.e., Marbe's Law) are highly correlated with commonality (Laffal, 1955; Moran et al., 1964), some information on these variables may be gleaned indirectly from observed variability in the commonality measure in the present study and in those to follow.

Prior to an account of the application of these lists in a series of studies, attention should be called to the relevance of idiodynamic associative sets to traditional commonality norm tables. Results on the present 482 students provide some information on the composition of word-association commonality norms in general.

Commonality Norms and Individual Set Hierarchies

In the early Würzberg studies, the commonality value of stimulus response word pairs ("individual strength of the reproduction"), interacting with momentary "determining tendency" ("operating task," or set) provided an account of "thinking." In these controlled association studies the subject usually was provided a series of different associative tasks (sets) by the experimenter. Watt's (1964) observations on the interaction between "strength of the reproduction" and the operation of sets provide an interesting and dynamic account of the results reported in Table 3, above.

> ...the influences which determine every event in our mental experience fall into two large groups, the operating task and the individual strength of the reproductions which come thereby into question. On the one hand, the task may find no reproductions, in which case no reaction can occur; and, on the other hand, the strength of the tendency to reproduction may be too great for the task to operate, in which case it forces its way out in spite of the task, or before any reproduction which the task favors has had time to become actual: in other words, a wrong reaction takes place. Otherwise more or less suitable reactions occur. This is thought to be *valid for the whole of our mental experience* ... [p. 194].

Watt, in setting different tasks (sets) for his subjects on the same stimulus words, observed directly the facilitative and inhibitive effects of set, depending upon compatibility of a specific set with an assumed a priori hierarchy of bond strength between word pairs. This concept of a universal, enduring, a priori hierarchy of associates of varying "strengths" as reflected by word-association commonality norms is widely held today, as indicated in a review of psy-

cholinguistics, by Rubenstein and Aborn (1960).

> ...a number of studies concerned with the probability of language segments and with word association have brought forth a point of view which stresses the significance of the concept of response hierarchy in interpreting the subject's performance in various verbal tasks. These studies have suggested, and to some extent supported, the following hypotheses:
>
> (a) Differential exposure to language segments (letters, words, etc.) produces in the individual a set of correlated probabilities of emitting those segments.
>
> (b) Since segments in natural languages are characterized by inequality in frequency of occurrence, experience with language—both in sending and receiving messages—imparts to the individual an isomorphic response hierarchy.
>
> (c) Because members of the same linguistic community share a common language experience, their response hierarchies are similar [p. 291].

The Minnesota norms may be used to illustrate the common-response hierarchy referred to above. The Minnesota commonality tables report the associates of about 1,000 students to the 100-word Kent-Rosanoff list. From these norms it would be determined, for example, that to WOMAN the word man has much greater association strength (given by 646 subjects) than has the word lady (given by only 15 subjects) and that to the word AFRAID, the word scared (240) has greater association strength than has the word brave (62) (Russell & Jenkins, 1954).

As Watt discovered, and as indicated in Table 3 and in Moran et al. (1964), the associates of a subject with a specific set frequently run counter to those predicted by such a common-response hierarchy. Unlike Watt's temporarily induced sets, the idiodynamic sets postulated in the present study are enduring characteristics, traits, of individual subjects. Since the associations of the three types of set deviate in a consistent, predictable manner from norms based upon the "generalized other" in the same linguistic community, it should be of interest to examine their deviate association hierarchies in relation to the general commonality norms. To this end, the associates of the subjects used to represent sets in Table 3 were tallied separately, providing independent commonality norms for object-referent, concept-referent, and speed-set subjects. The results are presented in Table 4.

On the left side of Table 4 are commonality norms on 482 subjects, presented in the traditional manner. Here it may be seen, for example, that the association strength of DIE-live is approximately twice that of DIE-death, and that of DIE-death about twice the association strength of DIE-dead, etc.

In the center of Table 4, the relative frequencies of "normative" primary, secondary, and tertiary associates are given separately for each of the three set types. Note that not one of the object-referent set subjects gave the "primary" associate to DIE or STEEL; not one of the concept-referent set subjects gave the primary associate to HIT or YELLOW; and virtually no speed-set subjects gave the primary associate to HAM or FIDDLE. These are extreme examples, to be sure. Examples like those given in Table 4 for PAIL and RADIO are much more common, however, and illustrate the same point: "Commonality itself is a complex variable, a resultant score derived from consensus within different set types [Moran et al., 1964, p. 10]." The effects of averaging across frequencies in the three individual response hierarchies may be seen on the far right of Table 4. These averages across sets yield essentially the same overall hierarchy as that of the commonality norms on the left in Table 4.

Later in this report, after an additional (perceptual-referent) set has been described and other aspects of associative sets examined, the arbitrariness of commonality norms as a measure of general "association strength" will be even more evident.

Idiodynamic Associative Sets of Spanish-Speaking Subjects in Collaboration with Rafael Núñez, National University of Mexico

Studies in which commonality tables of English-speaking subjects have been compared with similar tables based upon associates of foreign language subjects (e.g., Esper, 1918; Miron & Wolfe, 1964; Rosenzweig, 1964), have all found a great similarity in response hierarchies. The present study was undertaken to determine whether

TABLE 4
COMMONALITY AS AN AVERAGE OF INDIVIDUAL SET HIERARCHIES

Commonality table (%)			Set hierarchies (%)			
Stimulus	Response	N = 482 Commonality (%)	N = 47 Object-referent	N = 32 Concept-referent	N = 53 Speed	Average (%) of three sets
DIE						
	Live	27	0	12	72	28
	Death	12	11	28	7	15
	Dead	6	0	12	0	4
STEEL						
	Iron	18	0	9	28	12
	Thief	8	21	6	9	12
	Rob	8	0	16	9	8
HIT						
	Hurt	32	32	0	15	16
	Strike	14	13	31	15	20
	Run	10	11	9	9	10
YELLOW						
	Green	14	6	0	36	14
	Blue	8	4	0	19	8
	Color	8	6	22	0	9
HAM						
	Pig	20	43	22	4	23
	Eggs	17	11	6	36	18
	Meat	7	0	28	0	9
FIDDLE						
	Music	25	62	38	7	36
	Violin	21	6	28	30	21
	Play	9	17	9	0	9
PAIL						
	White	23	4	41	11	19
	Sick	13	21	16	0	12
	Bucket	12	8	0	23	10
RADIO						
	Music	35	68	25	13	35
	TV	29	6	16	74	32
	Listen	6	8	9	0	6

this interlinguistic similarity extended to idiodynamic sets.

Procedure

The subjects consisted of four groups of Spanish-speaking students, 206 in all, at the National University of Mexico. The 80-word list was read in Spanish, in the same manner as in the preceding study, and responses written in Spanish were scored clerically from a prescored manual, based upon the associates of the 482 English-speaking sample, with commonality scored in the same manner, but based upon the 482 sample instead of the original sample of 196 men.

Findings

The means, standard deviations, and intercorrelations of the variables are provided in Table 5. Shown in Table 6 are the results of a normalized varimax rotation of Principal Component factors.

The same three factors found in earlier samples of English-speaking subjects appeared also in the Spanish-speaking sample. In Table 6, contrast-coordinate (Factor I), synonym-superordinate (Factor II), and functional (Factor III) clearly represent the three independent associative modes.

To demonstrate idiodynamic sets, subjects were selected to represent a set if they had a standard score greater than .50 on the variables representative of a set and less than .50 on the variables representative of the other two sets, based upon their first 44 associates only. These cutting scores, rather than 1.00 versus .00 as in the preceding study, were used because of the smaller sample and, of course, yielded groups with less pronounced sets. The prediction was made that groups evidencing a set would

TABLE 5

Means, Standard Deviations, and Intercorrelations for 206 Mexican Students

Variable	Mean	SD	Intercorrelations 1	2	3	4	5	6
1. Functional	15.2	5.2						
2. Synonym	4.6	2.8	.04					
3. Superordinate	6.0	4.5	−.06	.36				
4. Contrast	5.9	5.1	.01	−.09	−.17			
5. Coordinate	5.7	4.7	−.19	.11	−.09	.64		
6. Faults	4.7	5.6	−.32	−.10	.10	−.25	−.19	
7. Commonality	4111.3	1615.4	.45	.17	.03	.75	.48	−.32

achieve a higher commonality score on the 12 criterion stimulus words in the second part of the lists that were most compatible with their set.

Because the means on commonality for the three different groups of 12 words were very different (functional, 1,163; synonym-superordinate, 380; contrast-coordinate, 1,525), standard scores (based upon a total sample of 206) were reported in Table 7. The reason for these mean differences is unclear; some of the differences may be attributable to difficulties in translation of both the stimulus and the response words. The important point illustrated by Table 7, however, is the indication of the operation of set. In every instance, the sample of subjects selected to represent a set achieved a higher commonality score on the criterion stimulus words most compatible with their set, as predicted. Clearly, the Mexican Spanish-speaking college students evidenced the same three idiodynamic associative sets observed in English-speaking subjects.

TABLE 6

Normalized Varimax Rotated Factors for 206 Mexican Students

Variable	Factor I	II	III	h^2
1. Functional	−.16	−.01	.91	85
2. Synonym	.07	.83	.11	70
3. Superordinate	−.10	.82	−.10	69
4. Contrast	.90	−.16	.16	87
5. Coordinate	.89	.04	−.12	81
6. Faults	−.24	.02	−.64	47
7. Commonality	.71	.14	.56	85

To this point, the same three idiodynamic sets have been demonstrated in a noncollege group of 35-year old normal men, an acutely psychotic group of schizophrenic patients (Moran et al., 1964), a group of college freshmen, and a group of Mexican college students. The "existence" of such sets seems to be well established in the most diverse groups of adults.

The five variables used to measure sets accounted for one-half to two-thirds of all associates to the present word list (Tables 5 and 1). Other kinds of idiodynamic sets may account for additional unscored associates. In fact, Jung's (1919) "predication type" is as yet unrepresented in the present system.

TABLE 7

Effect of Idiodynamic Set upon Commonality Score

N	Type of subject	Type of stimulus word: Functional	Synonym-superordinate	Contrast-coordinate
		(Average commonality standard score)		
32	(F) Object-referent	.31	−.23	−.25
34	(SS) Concept-referent	−.11	.52	−.55
35	(CC) Speed set	.04	.02	1.10

Note.—Type of subject represented by subjects with z score greater than .50 on variables representing one set and less than .50 on variables representing the other two sets, based upon first 44 associations. Type of stimulus word each represented by 12 words that had elicited predominately one type of response from 196 men; these were the last 36 stimulus words in the list.

A Perceptual-Referent Associative Set

One group among the normal subjects of Jung's (1919) word-association experiments evidenced a very strong tendency to give noun responses to adjective stimulus words.

> This preference for the noun arises from the endeavour of the predicate type to react chiefly by attributes. Our figures show, not merely that a predicate is thus reacted but also, inversely a noun is given to the adjective when this is the stimulus word [pp. 166–167].

Jung regarded the predicate type as a personality trait, observing that

> the predicate attitude is not accidental but corresponds to a definite psychological disposition, which is maintained even when other kinds of reactions would be much easier than the predicate forms [p. 167].

Siipola has investigated the predicate type (a-noun) in relationship to a set which she considered to be its polar opposite, the set to give contrast associates (Dunn, Bliss, & Siipola, 1958; Siipola, Walker, & Kolb, 1955). Siipola et al. (1955) attributed the two associative sets to differences in attitudes toward speed in responding.

> ...individuals adopt their own self-defined attitudes toward the speed at which they will operate, varying all the way from excessive concern with speed to feelings of almost complete freedom from time pressure.... High production of *contrast* responses and low production of *a-noun* responses are characteristic effects of an attitude of time pressure regardless of whether this attitude is experimentally imposed or is self-induced by the subject [p. 450].

Other investigators have reported a negative correlation between predication and contrast associates, for example, −.45 (Wells, 1912), −.55 (Kelley, 1913), −.69 (Tendler, 1933). In their Table 4, Moran et al. (1964) reported for contrast associates correlations of −.60 and −.56 with "intrinsic predicate" (e.g., RED-apple) and "extrinsic predicate" (e.g., ROTTEN-apple) associates, respectively. For the logical coordinates, comparable correlations were −.52 and −.58, respectively.

The tendency to give predominately predication associates seems to represent a fourth idiodynamic associative set, characterized also by slow reaction time and a marked tendency not to give contrast or coordinate associates. From the foregoing it should be predicted that a predication variable would have a large loading at the opposite pole of the contrast-coordinate factor.

Procedure

For purposes of reliability studies, to be reported later in this paper, a sample of 327 freshmen earlier had been administered the 80-word list in the manner described for the sample of 482 freshmen. This sample of 327 was scored by the manual based on 196 men (Moran et al., 1964), since the manual based on the 482 freshmen had not yet been constructed. A predication variable was added to those already scored.

Predication associate. The stimulus word and response word are adjective-noun or noun-adjective combinations; the stimulus word denotes an attribute of the object denoted by the response word, or vice versa, for example, RED-apple, or APPLE-red.

Findings

The means, standard deviations, and intercorrelations of the variables are provided in Table 8. Shown in Table 9 are the results of a normalized varimax rotation of principal component factors.

Factor I, in Table 9, represents the so-called "speed" factor, with contrast-coordinate associates and predication associates at opposite poles, as predicted by Siipola. Factors II and III are characterized by the functional and by the synonym-superordinate associates, respectively, with coordinate loading −.57 on II, and superordinate loading only .59 on III. Although this factor structure was not as sharply defined as those previously reported, the same three factors are unmistakable, with predication associates clearly negative to contrast-coordinate, and orthogonal to the other two factors.

Jung and Siipola both stressed the prominent role of imagery in predication associates. "As we have already suggested, the individuals who belong to the predicate type have, we assume, primarily vivid inner pictures ... [Jung, 1919, p. 160]"; the predication types "give more adjective-noun associates, and generally report complex processes, especially visual imagery, inter-

TABLE 8

MEANS, STANDARD DEVIATIONS, AND INTERCORRELATIONS FOR 327 COLLEGE FRESHMEN

Variable	Mean	SD	Intercorrelations						
			1	2	3	4	5	6	7
1. Predication	4.8	3.4							
2. Functional	17.3	5.2	.03						
3. Synonym	9.9	3.8	−.21	−.03					
4. Superordinate	4.9	2.6	−.24	−.02	.23				
5. Contrast	9.6	4.1	−.51	−.02	.00	.13			
6. Coordinate	10.4	4.5	−.41	−.36	.17	.03	.42		
7. Faults	6.5	5.3	−.02	−.20	−.34	−.25	−.29	−.27	
8. Commonality	1213.3	303.3	−.50	.35	.41	.37	.67	.33	−.44

vening between stimulus and response [Dunn, Bliss, & Siipola, 1958, p. 76]." The associates of predication-type subjects in the present study (20 subjects who were high on the predication end of Factor I and low on the other two factors) were overwhelmingly perceptual referent. Even on the last 36 stimulus words, which typically evoke one predominate associate, these subjects seldom gave the "typical" associate. The stimulus word NAIL had elicited from 52% of the normative group the object-referent associate, hammer. The present 20 predication-type subjects gave instead a variety of associates such as: broken, board, hard, steel, long, finger, window, cross, etc. Instead of giving the superordinate "bird" to CROW (as did 50% of the normative group), half of these subjects gave black. Instead of the popular response "sweet" to SOUR, predication-type subjects gave pickle (2), milk (2), lemon (3), orange, etc. Whereas the majority of normative subjects gave white to BLACK, these subjects gave associates like night (3), Negro, dress, Zorro, suit, smoke, etc.

TABLE 9

NORMALIZED VARIMAX ROTATED FACTORS FOR 327 COLLEGE FRESHMEN

Variable	Factor			h^2
	I	II	III	
1. Predication	.78	.13	−.08	62
2. Functional	−.04	.94	.08	88
3. Synonym	.00	−.16	.80	67
4. Superordinate	−.12	.00	.59	37
5. Contrast	−.89	.05	.06	79
6. Coordinate	−.59	−.57	.18	70
7. Faults	.16	−.17	−.71	56
8. Commonality	−.72	.32	.52	88

The "perceptual" quality of the associates is readily apparent. To stimulus words that denote objects, these subjects associated a perceptible (usually visual) attribute; to stimulus words that denote an attribute, these subjects associated an object which might display the attribute. Although one cannot predict from knowledge of a predication set the highly probable specific response word to a specific stimulus word (as often can be done for object-referent, concept-referent, and speed-set subjects), one can predict that the stimulus-response word pair will be descriptive of some object-attribute relationship. This consistent associative tendency might be termed a *perceptual-referent* set.

In addition to adjective-noun, noun-adjective associates and a tendency to report visual imagery, the perceptual-referent set subjects are characterized by very low commonality scores. Jung (1919) commented on the large number of "egocentric" associates given by such subjects. Wells (1912) reports a correlation of −.74 and Tendler (1933) one of −.79, between commonality and predication associates; the comparable correlation in the present study was −.50 (Table 8). The reason for this typically low commonality score is readily apparent from the scattered variety of object-attribute associates, described above, given by perceptual-referent set subjects.

To characterize the tendency to give

predicate associates as an idiodynamic perceptual-referent set implies an enduring personality trait. Jung (1919) conceptualized the predication-type in this manner:

> From the figures of the distraction experiment it can be stated that the predicate type is no merely accidental momentary attitude, but corresponds to an important psychological characteristic—one which is maintained amid altered conditions [pp. 162–163].

Wells (1912) also has reported a correlation of .83 for predication associates, in subjects retested after a 14-month interval. Reliability is a critical issue for dispositional constructs, such as the postulated idiodynamic associative sets, and will be examined next.

Reliability Studies

The present scoring system consists of six variables, used to represent four postulated idiodynamic associative sets, plus a fault and a commonality variable. Such a system categorized about 80% of all associations (Table 8). In passing, it is interesting to note that Woodworth (1948) long ago intuitively derived a scoring system much like this one. Woodworth remarked,

> There does seem to be some psychological basis for the fourfold classification suggested; but it is doubtful whether any such scheme can win general acceptance at the present time [p. 353].

An examination of the reliability of this system follows.

Factor Invariance: Replication Studies

The three orthogonal factors, represented by synonym-superordinate, contrast-coordinate, and the functional variable, appear to be exceptionally stable with different word lists (Moran et al., 1964) and across quite different subject populations. Introduction of a predication variable in the preceding study yielded a bipole predication versus contrast-coordinate factor.

In order to determine the replicability of this "new" factor structure, with the additional predication variable, a different sample of 353 freshmen was tested with the 80-word list. The procedures followed with the new sample of 353 freshmen, except for use of the scoring manual based upon 482 freshmen (Table 1) instead of the 196 men, were identical to that of the preceding study. Means, standard deviations, and intercorrelations are provided in Table 10. Table 11 shows the results of a normalized varimax rotation of principal component factors.

Comparison of the factor loadings in Tables 9 and 11 reveals essentially the same

TABLE 11

Normalized Varimax Rotated Factors for 353 College Freshmen

Variable	Factor I	Factor II	Factor III	h^2
1. Predication	.68	.06	−.27	53
2. Functional	.24	.90	−.01	87
3. Synonym	−.14	−.07	.84	73
4. Superordinate	.02	.20	.76	62
5. Contrast	−.89	.10	−.14	82
6. Coordinate	−.86	−.27	.00	81
7. Faults	.44	−.50	−.21	49
8. Commonality	−.69	.48	.33	82

TABLE 10

Means, Standard Deviations, and Intercorrelations for 353 College Freshmen

Variable	Mean	*SD*	1	2	3	4	5	6	7
1. Predication	5.3	3.2							
2. Functional	17.0	5.1	.05						
3. Synonym	9.8	3.9	−.23	−.08					
4. Superordinate	5.3	2.8	−.21	.14	.31				
5. Contrast	8.3	3.9	−.53	−.15	−.62	−.02			
6. Coordinate	9.7	4.6	−.50	−.42	.13	−.07	.61		
7. Faults	6.1	4.9	.07	−.19	−.28	−.16	−.32	−.33	
8. Commonality	8065.7	1935.1	−.45	.26	.38	.29	.62	.44	−.46

TABLE 12

NORMALIZED VARIMAX ROTATED FACTORS FOR 206 MEXICAN COLLEGE STUDENTS

Variable	Factor I	Factor II	Factor III	h^2
1. Predication	.67	.31	−.42	72
2. Functional	.13	.84	.07	73
3. Synonym	−.04	.16	.77	62
4. Superordinate	.05	−.13	.83	71
5. Contrast	−.89	.20	−.19	87
6. Coordinate	−.86	−.05	−.02	74
7. Faults	.15	−.72	.08	54
8. Commonality	−.71	.57	.14	85

factors in both analyses, with coordinate this time having a lower negative loading (−.27) on the "functional" factor in Table 11. These results on a different sample of 353 subjects indicate the replicability of factor structure, with the inclusion of the new predication variable.

This replicability is further demonstrated in a reanalysis of the associates of the 206 Spanish-speaking subjects, first shown in Table 6. For the present reanalysis, associates were scored also for predication responses, and refactored, with results as shown in Table 12. Even when the word-association experiment was conducted entirely in Spanish, the three factors appeared, with the predication variable again principally on the opposite end of the contrast-coordinate factor.

FACTOR INVARIANCE: MEN VERSUS WOMEN

The 327 freshman sample (Tables 8 and 9) was divided into two samples of 173 men and 154 women, and each sample was factor analyzed separately. Two sets of factor scores were calculated for the men; one set of factor scores was based upon beta weights derived from the factor analysis on the men, but the other set of factor scores, for the same men, was based upon the beta weights derived from the independent factor analysis on the women. Correlations between the two sets of factor scores were: I. perceptual-referent and speed-set factor, .94; II. object-referent set factor, .98; III. concept-referent set factor, .98. These correlation coefficients are of the same order as the correlations obtained when the same sample is tested on two occasions, as described below. Clearly, sex of the subjects did not affect factor structure.

FACTOR INVARIANCE: TEST-RETEST

Three months after the first testing, the same word list was administered, in reverse order, to 195 of the 372 freshman sample. Results of first testing and second testing were factored separately for the 195 subjects. Two sets of factor scores were calculated for the first test performance only. One set of factor scores on the first performance was based upon the beta weights derived from a factor analysis of the first performance. Another set of factor scores on the same first performance was based upon the beta weights derived from the factor analysis of the later, second performance. Correlations between the two sets of factor scores were: I. perceptual-referent and speed-set factor, .97; II. object-referent set factor, .96; III. concept-referent set factor, .99. Thus, the factor structures were so similar on the two occasions, 90 days apart, that it actually made little difference which structure was used to calculate factor scores.

Interrelationships among the eight variables discussed so far appear to be best represented, at least for adult subjects, by the three orthogonal factors described above. Factor structure, per se, seems to be highly reliable. The next step, however, involves an attempt to better understand specific associative tendencies of individual subjects, guided by the score patterns on the three factors. It thus becomes important to determine the reliability of subjects, with respect to scores on these factors.

Reliability of Factor Scores

To determine the split-half reliability of factor scores, associates to List A and to List B were factored separately for the sample of 327 freshmen. Factor scores derived from the two different factor structures were then correlated, and corrected by the Spearman-Brown formula for an 80-word list. The two sets of factor scores on I. perceptual-referent and speed set correlated .84; on II. object-referent set, .69; and III. concept-referent set, .83. Thus, subjects

TABLE 13

Rotated Orthogonal Factor Loadings of Factor Scores for 195 Subjects

Test session	Variable	Factor	I	II	III	h^2
First	1	I (Perceptual-speed)	.94	−.01	−.01	89
First	2	II (Object-referent)	.03	.92	.03	84
First	3	III (Concept-referent)	−.01	−.05	.89	80
(90-Day interval)						
Second	4	I (Perceptual-speed)	.94	.03	−.05	88
Second	5	II (Object-referent)	−.01	.92	−.02	85
Second	6	III (Concept-referent)	−.05	.06	.89	79

tended to maintain relative rank positions on the three dimensions represented by the three factors, on both word lists.

Three months after the first testing session, 195 of the 327 sample had been retested. These data were used above in the examination of factor structure. In the present analysis, factor scores were calculated separately from factor analyses of each separate testing, to determine whether subjects maintained relative rank position on two different occasions, that is, test-retest reliability of subjects on the three factors. The six factor scores, three from each testing, were treated as variables and factor analyzed jointly, with results as shown in Table 13. It may be seen that the three factors were almost perfectly orthogonal on both occasions. Since individual subject factor scores were employed, the similarity of factor loadings on both occasions reflects directly the test-retest consistency of the subjects. Correlations between factor scores derived separately from first and second test factor analyses were: I. perceptual-referent and speed set, .75; II. object-referent set, .65; III. concept-referent set, .58.

Reliability of Individual Variables

A comparison of performances on Lists A and B, based upon the 327 sample, is provided in Table 14. Means and standard deviations on the eight variables were roughly comparable for the two lists. Considering the low mean frequency of some associates, the split-half reliability coefficients for the eight variables indicate reasonable consistency of subjects on the two lists.

When the same word list was readministered (in reverse sequence) to a sample of 195 subjects after a 90-day interval the only appreciable mean difference was a lower number of response faults and a slightly higher commonality score, on second testing, as shown in Table 15. A general, but very slight, drop in standard deviations on second testing also may be seen in Table 15. The test-retest correlations on faults was low (.49) and for superordinate, very low (.36). Otherwise, the test-retest correlations

TABLE 14

Means, Standard Deviations, and Intercorrelation of List A and List B, for 327 College Freshmen

Variable	List A		List B		r_{tt}
	Mean	*SD*	Mean	*SD*	
1. Predication	2.3	1.9	2.5	1.9	.77
2. Functional	8.9	3.0	8.3	3.1	.61
3. Synonym	5.3	2.2	4.6	2.1	.67
4. Superordinate	2.2	1.6	2.7	1.6	.56
5. Contrast	4.7	2.2	4.9	2.4	.79
6. Coordinate	5.3	2.5	5.0	2.6	.73
7. Faults	3.0	2.6	3.5	3.1	.82
8. Commonality	627.8	165.8	585.5	162.1	.83

Note.—r_{tt} = Pearson product-moment correlation coefficient, corrected by Spearman-Brown formula for total test, 80 words.

TABLE 15

Means, Standard Deviations, and Intercorrelation of 90-Day Test-Retest Scores of 195 College Freshmen

Variable	1st Test		2nd Test		r
	Mean	SD	Mean	SD	
1. Predication	4.9	3.5	4.2	3.3	.68
2. Functional	17.4	5.2	16.7	4.7	.66
3. Synonym	10.0	3.8	10.1	3.4	.62
4. Superordinate	4.8	2.7	5.2	2.3	.36
5. Contrast	10.0	3.9	10.3	3.8	.71
6. Coordinate	10.5	4.6	11.4	4.3	.63
7. Faults	6.1	4.9	4.0	3.7	.49
8. Commonality	1233.3	263.9	1293.2	266.1	.70

of individual variables, shown in Table 15, indicate moderate consistency in performances 90 days apart, the coefficients ranging from .62 to .71.

In the studies that have been described herein, and in Moran et al. (1964), replicability of the idiodynamic associative set phenomena seems to have been reasonably well established. By following the same, or similar, "operations" as outlined in these studies, other investigators ought to be able to reproduce these set phenomena at will. In what follows, task attitudes (Siipola et al., 1955) are investigated in relation to the individual differences in linguistic habits which give rise to idiodynamic associative sets.

Task Attitudes and Idiodynamic Sets

Siipola has demonstrated what appears to be a reciprocal relationship between predication and contrast associates, varying as a function of the subject's attitude toward speed in responding. For example, she has shown that when subjects were placed under greater time pressure the number of contrast associates increased markedly (from 174 to 442) and the number of adjective-noun associates decreased (from 705 to 392) just as markedly (Siipola et al., 1955). The bipole Factor I in Tables 9, 11, and 12, with predication at one end and contrast-coordinate at the other, is compatible with Siipola's view. It need not follow from these findings, however, that perceptual-referent-set subjects shift to give contrast associates under time pressure, or that speed-set subjects shift to give predication associates when time pressure is lifted. It will be recalled that Siipola's word list was "loaded" with popular contrast-evoking stimulus words. As indicated earlier, it may be that time pressure moved people to give more popular associates and that the "category" into which such associates fall is determined by the experimenter's selection of stimulus words. Also, even if time pressure did induce people in general to shift "category" of associate, those people with strong idiodynamic sets might not do so. In fact, Jung (1919) found his predicate-type subjects to be practically impervious to various forms of distraction.

> For the sake of comparison we have placed by the side of the predicate type the average of all other types. The difference is striking. In distraction the predicate type shows no change worth mentioning. The predicate type does not dissociate its attention, whilst all other types prove themselves, to some extent at least, accessible to the disturbing stimulus. This fact is extremely remarkable [p. 160].

Jung (1919) stressed the prominent role of imagery in predicate-type subjects.

> The predicate type is unable to dissociate his attention because his primary vivid inner pictures make such a demand upon his attention that inferior associations (which make up the distraction phenomenon) cannot arise [p. 162].

Siipola also observed that such subjects "generally report complex processes, especially visual imagery, intervening between stimulus and response [Dunn et al., 1958, p. 76]."

The present experiment was designed to examine the effects of time pressure upon

idiodynamic associative sets and upon imagery.

Procedure

Two months after the first testing, 240 of the 353 sample (Tables 10 and 11) were retested in the following manner. Subjects were instructed, as before, to write down the first word that came to mind when they heard the stimulus word. They were told that the words would be read in rapid order, and they would have to write quickly. If no word came to mind, they were to leave the space blank. The 40 words in List A were then read at 4-second intervals. The subjects then were told that stimulus words would be read very slowly and that they would have a great deal of time to write down the word that came to mind when they heard the stimulus word. If no word came to mind they were to leave the space blank. The 40 words in List B were then read at 8-second intervals. After both lists had been administered, the subjects were asked to review their first 40 associates and to estimate the number of times an associate had been accompanied by an image (a mental picture). Several minutes were allowed for the review, then subjects were asked to record the number of images at the top of the page. The same procedure was followed to obtain number of images on the second 40 associates.

Later, a comparable freshman class of 246 students, taught by the same instructor, was administered the two-word lists by the same examiner. Procedure was identical to that just described, except that List B was read first at 4-second intervals, followed by list A at 8-second intervals. To equate the two samples, 6 subjects were discarded from the 246 sample.

In the sample of 353, tested 2 months earlier, subjects had been asked to record the number of images experienced on the 80-word list. These data also were brought to bear in the present analysis.

Findings on the Total Sample

Overall mean differences under the 4-second and 8-second condition are shown in Table 16. Also provided are the means of a sample of 353, which had associated to both List A and List B at 5-second intervals. This comparison sample is useful in differentiating initial list differences from differences due to testing condition. The statistical significance values reported, however, are based upon the total 480 sample, for whom the lists were counterbalanced.

Mean differences under the two testing conditions were generally very slight, as may be seen in Table 16, but the results corroborate the findings of Siipola et al. (1955). The frequency of contrast-coordinate associates was significantly higher, and the incidence of predicate associates and reported imagery was significantly lower under the 4-second, time-pressure condition. A slight increase in commonality and decrease in synonym associates under time pressure also were statistically significant.

The most impressive effect of time pressure was that upon reported imagery. Of the 194 in the first sample who reported a difference in incidence of imagery under the two conditions, 146 reported a decrease under the 4-second condition; of the comparable 212 in the second sample (with word

TABLE 16

Mean Changes under 4- and 8-Second Conditions

	Comparison sample $N = 353$[a]			First sample $N = 240$		Second sample $N = 240$		Total sample $N = 480$		p[b]
	List B 5 seconds	List A 5 seconds	Lists A + B 5 seconds	List B 8 seconds	List A 4 seconds	List A 8 seconds	List B 4 seconds	Lists A + B 8 seconds	Lists A + B 4 seconds	
Predication	2.7	2.6	5.3	3.1	2.1	2.4	2.2	5.4	4.3	.01
Functional	8.3	8.7	17.0	8.0	8.4	8.2	7.9	16.2	16.3	n.s.
Synonym	5.0	4.8	9.8	5.0	4.8	5.3	4.7	10.2	9.5	.02
Superordinate	2.8	2.5	5.3	3.1	2.6	2.7	3.2	5.8	5.8	n.s.
Contrast	4.3	3.9	8.2	4.1	4.0	4.6	5.2	8.7	9.1	.05
Coordinate	4.4	5.2	9.6	4.8	5.4	5.2	5.4	10.0	10.8	.01
Faults	3.2	2.9	6.1	2.3	2.7	1.1	1.4	3.4	4.1	.01
Images			36.8	20.2	16.5	21.8	18.7	42.0	35.2	.01
Commonality	3657	4409	8066	3623	4622	4348	3954	7971	8576	.02

[a] See Table 10.

[b] Significance level determined by two-cell chi-square, comparing observed frequency of increased versus decreased scores under 4- and 8-second condition, with expected frequency of 50–50 (%); significance levels are for two-tailed test, $N = 480$.

lists reversed), 164 reported a decrease under the 4-second condition. The incidence of reported imagery on List A correlated .71 with that reported on List B. Two-month test-retest correlations between reported imagery on the 80-word list read at 5-second intervals and that on the 40-word lists read at 4-second and 8-second intervals (first 240 sample) were .43 and .38, respectively.

The very dramatic shift from predicate to contrast associates under time pressure, as reported in the Siipola et al. (1955) study, was much less dramatic when repeated with word lists that were not loaded with contrast-evoking stimulus words. Nevertheless, the associative shifts of people in general were as predicted by Siipola. It will be of interest now to examine the effects of time pressure upon idiodynamic associative sets.

Findings on Idiodynamic Sets

Results on the two 240 samples were factored separately for the 4-second and 8-second condition, yielding four independent factor analyses. There were no differences in factor structure attributable to differences in testing conditions; the same three factors appeared in all analyses. To determine effects of time pressure, subjects were selected to represent a set by their factor scores on the 8-second condition and then the factor scores on the 4-second condition were examined. Results on the two 240 samples are shown separately in Table 17, to provide a replication with word lists reversed.

Perceptual-referent set. The 36 subjects in the first 240-subject sample who had high scores (>.50) on Factor I and low scores (<.30) on Factors II and III in the 8-second condition, showed virtually no change in factor scores based upon the 4-second condition (Table 17). The stability of the set to give predicate associates was, in the words of Jung, "extremely remarkable." Results on the 30 subjects in the second 240-subject sample, with lists reversed, were somewhat less consistent. Despite the tendency of this group to shift to more "functional" associates (.30), however, the set to give predicate associates remained clearly predominate (.96).

Examination of score changes for the total 65 perceptual-referent set subjects on the individual variables showed, under the 4-second condition, a decrease in predicate associates and reported imagery, with a concomitant increase in associates of the other "set" categories, and in commonality. As shortly will be seen, these changes are typical of the other idiodynamic sets as well: a decrease in set-representative associates, with an increase in popular associates of the other categories. It would appear

TABLE 17

FACTOR SCORE CHANGES UNDER 4- AND 8-SECOND CONDITIONS

Factor and variable	N	Factor scores of subjects with set: From first sample of 240, List B, 8 seconds			List A, 4 seconds			N	From second sample of 240, List A, 8 seconds			List B, 4 seconds		
		I	II	III	I	II	III		I	II	III	I	II	III
I Predication (+)	36	1.19	−.65	−.68	1.03	−.60	−.58	30	1.29	−.72	−.79	.96	.30	−.33
I Contrast-coordinate (−)	28	−1.28	−.69	−.70	−.82	−.66	.03	35	−1.06	−.59	−.52	−.56	−.43	.25
II Functional	12	.02	1.01	−.64	−.08	−.04	.13	13	.01	1.32	−.49	−.24	.09	−.04
III Synonym-superordinate	15	−.02	−.49	1.04	.17	−.03	.14	17	−.01	−.33	1.16	−.23	.04	.37

Note.—Set types represented by subjects with factor scores distributed as follows (based upon 8-second-condition factor structure): I. Predication, >.50 on I, II and III <.30; I. Contrast-coordinate, <−.50 on I, II and III <.30; II. Functional, >.50 on II, III <.30, I between .30 and −.30; III. Synonym-superordinate, III >.50, II <.30, I between .30 and −.30.

from the evidence in Table 17 that the perceptual-referent set was the least disturbed by time-pressure instructions.

Speed set. The 28 subjects chosen from the first 240 sample to represent the speed set (contrast-coordinate) under the 8-second condition, maintained the same set under the 4-second condition (Table 17). The decrease in set-representative associates (from −1.28 to −.82) was larger than that found with perceptual-referent set subjects, and a sharp increase in synonym-superordinate associates (from −.70 to .03) was notable. But factor scores under the 4-second condition remained well within the limits specified for speed-set subjects. Results on 35 subjects, chosen in the same manner from the second sample of 240, were about the same, with a somewhat larger decrease in set-representative associates (from −1.06 to −.56) and an increase in synonym-superordinate associates (from −.52 to .25). Overall, it may be said that speed set was "weaker," but still much in evidence under the 4-second condition.

Inspection of score changes on the individual variables for the total 63 speed-set subjects revealed the same pattern of shifting as that noted in the perceptual-referent set subjects. In the 4-second condition, the frequency of set-representative contrast-coordinate associates dropped from 14.2 to 11.1, frequency of associates in the other categories increased (except predication), with an increase in commonality, and a decrease in reported imagery.

The speed set was so called mainly because of evidence that when people in general were placed under time pressure in the word-association experiment, the frequency of contrast-coordinate associates markedly increased (Siipola et al., 1955). This phenomenon was demonstrated also, to a lesser degree, in Table 16. Contrast-coordinate associates also tend to have fast reaction time (Moran et al., 1964). On this basis, it was inferred that some subjects had their own enduring, built-in set for speed and, for this reason, consistently behaved as people in general do when under explicit time-pressure instructions: that is, they produce many contrast-coordinate associates.

The data presented above are incompatible with such a rationale. The linguistic habit which underlies the set to give contrast-coordinate associates was disrupted, rather than facilitated, by explicit time-pressure instructions. It does not seem likely that these subjects initially were set for speed in responding, and therefore tended, only incidentally, to give many fast reaction-time contrast-coordinate associates (Moran et al., 1964). Rather, it would appear that this idiodynamic associative set promotes, specifically, contrast-coordinate associates. The enduring attitude which might best account for this set is more likely one toward words in general, rather than one involved with rate of responding in the word-association experiment. Since both contrasts (bipole) and coordinates are dimensional terms, this set might be termed a dimension-referent set. The aptness of this name may become more apparent as data on this set accumulate later in this report.

Object-referent and concept-referent sets. The 12 subjects in the first 240-subject sample who were chosen by 8-second condition factor scores to represent the object-referent set showed no set at all in the 4-second condition. This finding was repeated with 13 subjects from the second 240-subject sample. Essentially the same result was obtained with the 32 concept-referent subjects, although some vestige of set remained (set factor score of .37) in the 4-second condition, with subjects drawn from the second 240-subject sample (Table 17).

Time pressure seemed to have the effect of dissolving object-referent and concept-referent associative sets. For example, under the 4-second condition concept-referent–set subjects gave about the same number of "functional" associates as did the object-referent subjects, the two set types gave superordinate associates with equal frequency, and both set types gave even more coordinate associates than the dimension-referent–set subjects. Considering the 90-day test-retest reliability of object-referent (.65) and concept-referent (.58) sets under the 5-second condition, the effect of time

pressure, as shown in Table 17, was quite marked. At present, little may be added to this descriptive account.

Although the effects of time pressure upon different idiodynamic sets varied greatly in degree, all four sets evidenced the same general change pattern under the 4-second condition: (*a*) a decrease in frequency of the category of associates representative of the set, (*b*) an increase in frequency of associates representative of the other sets, except predication, (*c*) an increase in commonality, (*d*) a decrease in frequency of predicate associates, and (*e*) a decrease in incidence of reported imagery.

The 186 subjects who participated in the above analysis of effects of speed instructions upon sets have been treated, to this point, as the 39% of "deviate" cases in the sample of 480. This treatment was, of course, quite arbitrary. Any number of subjects could have been included in the "deviate" or "nondeviate" samples simply by changing cutting scores. But the purpose of this series of studies has been to determine the generality, reliability, and characteristic features of specific idiodynamic sets, and the selection of "extreme" cases by means of arbitrary cutting scores has helped to serve this purpose. Findings on these subjects may now be generalized to the other subjects.

Bases for Matching Word Pairs

The high reliability of three orthogonal word-association factor scores on 4 successive days, with different stimulus words each day, led to the postulation of three idiodynamic associative sets (Moran et al., 1964). Obviously, persons consistently high on only one of the three factors had to be evidencing a strong tendency, or set, to give predominately those categories of associate which loaded highest on that factor. Subsequent studies were centered upon only those persons who evidenced a strong set. But the three-factor structure, which has appeared with great fidelity in quite different kinds of subject populations, represents the orthogonal dimensions which most parsimoniously account for variabilities in the multiple measures taken upon the *total* sample.

Factors as Generic Dimensions

To illustrate, what would happen to the three-factor structure if all subjects with sets (as defined in preceding studies) were removed from the sample, and results on the residual subjects factored? The results of such an analysis may be seen in Table 18. The 144 subjects with idiodynamic sets were removed from the 353 sample (Table 10), and the results of the residual 209 subjects were factor analyzed.

As indicated by the results in Table 18, removal of set "types" had little effect upon factor structure. Using the same cutting scores, all set types were then removed from the sample of 209 and the residual factor analyzed, again yielding the same three factors. This process was repeated until 223 subjects identified as types had been removed, and the residual sample of 130 was factor analyzed. The end product of this iterative process is presented in Table 19.

As shown in Table 19, removal of the deviate subjects did not have an appreciable effect upon factor structure. An analogy may be made here to the Thurstone (1957) "box problem." Multiple linear measurements taken on a sample of boxes will yield

TABLE 18

Normalized Varimax Rotated Factors of 353 Sample, after Removal of 144 Subjects with Sets, $N = 209$

Variable	Factor I	Factor II	Factor III	h^2
1. Predication	.68	.10	−.15	49
2. Functional	.18	.88	−.08	82
3. Synonym	−.20	−.20	.82	76
4. Superordinate	.10	.32	.70	60
5. Contrast	−.89	.15	−.11	82
6. Coordinate	−.84	−.26	.02	78
7. Faults	.36	−.53	−.29	50
8. Commonality	−.77	.42	.27	84

Note.—Subjects with sets were selected as follows: Perceptual-referent set—Factor I score >.50, II and III <.30; dimension-referent set—Factor I score <−.50, II and III <.30; object-referent set—Factor II score >.50, III <.30, I between .30 and −.30; concept-referent set—Factor II score >.50, II <.30, I between .30 and −.30.

TABLE 19

Normalized Varimax Rotated Factors of 353 Sample after Removal of 223 Subjects with Sets, $N = 130$

Variable	Factor I	Factor II	Factor III	h^2
1. Predication	.67	.11	−.24	52
2. Functional	.26	.82	−.14	76
3. Synonym	−.19	−.16	.82	74
4. Superordinate	.14	.33	.74	67
5. Contrast	−.86	.18	−.17	80
6. Coordinate	−.85	−.19	−.03	76
7. Faults	.21	−.67	−.17	53
8. Commonality	−.71	.50	.25	82

Note.—Subjects with "sets" were selected as follows: Perceptual-referent set—Factor I score >.50, II and III <.30; dimension-referent set—Factor score on I <−.50, II and III <.30; object-referent set—Factor II score >.50, III <.30, I between .30 and −.30; concept-referent set—Factor III score >.50, II <.30, I between .30 and −.30.

a three-factor solution: I, length; II, width; and III, height. If every box characterized by arbitrary cutting points as high on one factor but low on the other two factors was removed, a factor analysis of the "residual" sample of boxes would still yield the same three factors. This process could be repeated indefinitely. Analogously, the three word-association factors represent more generic dimensions of a domain partially tapped in a variety of different ways by the multiple individual variables employed in the preceding experiments.

Factors Interpreted as Bases for Matching Word Pairs

The three factors, or dimensions, may be interpreted as representing basic ways in which one word may be matched with another word. One argument against such a simplified account as this might be anticipated. Many very elaborate categorization systems for word associations have been published which demonstrate the seemingly infinite number of ways in which word pairs may be matched. Murphy (1917) for example, published an 87-category system for sorting word pairs in the free association experiment, with the comment: "I shall first give it as it stands entire, and then explain it, in so far as a thing so fearfully and wonderfully made can be explained [p. 248]." While numberless principles for pairing words may be conceived, however, the empirically derived factors in the preceding studies suggest a parsimonious set of word-matching principles to which most people seem to conform.

1. Words may be matched on the basis of object-attribute relationships. The experimenter names an attribute (e.g., BLUE) and the subject names an object (e.g., water), or vice versa (e.g., WATER, blue). Since the completed stimulus-response word pair always refers to a perceptible quality of a concrete thing, whether the experimenter names an attribute or names an object, these might be called perceptual-referent associates.

2. Words may be matched on the basis of the concrete things they name. The experimenter names a thing, and the subject names another thing associated with it, for example, SPIDER, web; FOOT, shoe; BOAT, dock. These might be termed object-referent associates: that is, the stimulus word is referred to the concrete object which it labels, and the response word is the label of another concrete object.

3. Words may be matched lexigraphically. The experimenter give a word and the subject "defines" it. The subject plays "dictionary," and gives a synonym or superordinate of the stimulus word: (stimulus word) implied "is" or "is a" (response word). Because the concrete things or actual events denoted are irrelevant to the "logical" basis of word matching used and since one concept is referred to another concept, these might be called concept-referent associates.

4. Words may be matched on the basis of a common dimension. The experimenter names one pole of a continuum (e.g., FAST), and the subject names the other pole (e.g., slow); or, the experimenter names one entity (e.g., BLUE), and the subject names another entity of the same logical order (e.g., green, or yellow, or pink, etc.). Since the stimulus-response word pair specifies or refers to a dimension, these might be called dimension-referent associates.

In all of the samples to which the present scoring system has been applied, the above word-matching bases have appeared in three orthogonal factors, with 1 and 4 as one bipole factor. One might expect this factor structure to be relatively unaffected by the language in which the word-association experiment is conducted (as with the Spanish-speaking Mexican subjects shown in Table 12), since the relationships involved are expressed in all languages. Subjects ordinarily utilize all four bases for matching words. Most subjects, however, also tend to use one basis more than other bases. A subject might consistently give concept-referent associates, for example, to the stimulus words used above to illustrate other bases for word matching, for example, SPIDER, insect; BLUE, color; FAST, rapid. It is this tendency to use one word-matching basis predominately and consistently that has been termed an idiodynamic associative set.

A Hierarchy of Word-Matching Bases

It is interesting to compare the notion of bases for matching words with that of bases for matching actual objects, which has been investigated extensively. Of the object-sorting behavior of adults, for example, it is reported that concrete (perception-bound), functional, and abstract (categorical) bases show, in that order, an increasing positive relationship to intelligence and educational level (McGaughran & Moran, 1956). Developmentally, the three object-matching bases also emerge in the above order (Reichard, Schneider, & Rapaport, 1944). Bruner & Olver (1965) have generalized these three matching bases to a developmental sequence in "modes of analyzing events":

> In short, then, the functional mode of analyzing events seems to develop before there is a full development of the superordinate strategies, and one is tempted to speculate that the shift from the consideration of surface, perceptible properties to more embracing functional properties may be the vehicle that makes possible the development of efficient and simpler grouping strategies [p. 433].

A parallel development in word-word matching bases would place the idiodynamic associative sets in the sequence: perceptual referent, object referent, concept referent.

The concept of "oppositeness," which underlies the dimension-referent matching basis, apparently develops slowly as a "mode for analyzing events." Kreezer and Dallenbach (1929) undertook to train 100 children, aged 5.0 to 7.5 years, to give opposites to 25 words. Children aged 5.0–5.5 could give, with training, only 40% correct opposites; children aged 6.0–6.5, 75% correct opposites; children aged 7.0–7.5, 90% correct opposites. Rabinowitz (1956), who used the Kreezer and Dallenbach training technique with 50 children, reported a correlation of .61 between IQ and ability to learn to give opposites. The dimension-referent associate seems to involve a fairly sophisticated matching basis.

In the light of the foregoing, the pattern of correlations shown in Table 20 is especially interesting. The correlations of the predicate variable with each of the other set-representative variables are given for several samples of subjects. A parallel account of Piaget's four periods in the development of thought, as summarized succinctly by Carroll (1964), is provided on the right side of the table.

As indicated in Table 20, each mode has a progressively greater negative correlation with predicate associates. The reliability of this phenomenon, with quite different kinds of subjects, word lists, and testing conditions, is substantial. The parallel between these correlation patterns and the ordered relationship of these four modes with respect to intelligence and developmental sequence suggest a hierarchical ranking in terms of linguistic sophistication.[3] Analogies with object-sorting behavior and coincidence with Piaget's account of the development of thought provide perhaps not the most convincing empirical evidence for

[3] For future reference, it might be noted here that the sum of unscored associates, treated as a single variable, correlated .32 with predication, −.40 with contrast, and −.31 with coordinate ($N = 353$); hence, such "heterogeneous" unclassified associates would appear at the base of this hierarchy. This "variable" correlated .48 with Factor I, Table 11 ($N = 353$), and .64 on test-retest (90-day interval, $N = 195$).

TABLE 20

CORRELATION OF OTHER VARIABLES WITH THE PREDICATION VARIABLE

	Subject samples									
	A College	B College	C Mexican	D College	E College	F College	G College	H Veterans	I Schizophrenic	Piaget's four periods in the development of thought
Predicate associates										(Age 2) "He has to learn to perceive certain aspects of his environment as invariant . . . these *perceptual invariants* may be thought of as the basis of thought and language."
Functional versus predicate	.03	.05	.16	−.06	.01	.03	.01	.17	.30	(Age 2–7) "In the next stage . . . the child wrestles with . . . interpretation of his environment, namely, the understanding of *relationships among the perceptual invariants*."
Synonym versus predicate	−.21	−.22	−.18	−.28	−.18	−.12	−.15	−.28	−.13	(Age 7–11) " . . . he can perform such operations as *substitution* and the *recognition of equivalences* . . ." " . . . passes into the stage of concrete operational thinking . . . he can *classify* objects into groups."
Superordinate versus predicate	−.24	−.22	−.30	−.19	−.20	−.17	−.13	−.22	−.05	
Coordinate versus predicate	−.41	−.50	−.47	−.40	−.47	−.51	−.53	−.44	−.19	(Age 11 on) " . . . starts to think in terms of *purely logical propositions* . . . some concepts may be built out of other concepts. Take the concept of "oppositeness," which must be built out of instances in which it is noticed that one extreme of any dimension of sensation is *contrasted* with the other extreme." (Carroll, 1964, pp. 78–81, italics added)
Contrast versus predicate	−.51	−.53	−.47	−.44	−.52	−.46	−.51	−.42	−.22	

Note.—Sample A, 327 college students (Table 8); B, 353 college students (Table 10); C 206 Mexican students; D, first 240 student sample, 4-second condition; E, same 240, 8-second condition; F, second 240 student sample, 4-second condition; G, same 240, 8-second condition; H and I, 79 normal older veterans and 79 schizophrenics, respectively, from Tables 4 and 8 in Moran et al. (1964).

ranking perceptual-referent, object-referent, concept-referent, and dimension-referent word-matching bases in a hierarchy of increasing linguistic sophistication. Some support for such a view may be found, however, from several sources.

Vygotsky (1962) has described the development of linguistic sophistication as an increasing consciousness of language itself as a complex system. Linguistic sophistication

> presupposes a hierarchy of concepts of different levels of generality. Thus the given concept is placed within a system of relationships of generality. The following example may illustrate the function of varying degrees of generality in the emergence of a system: A child learns the word *flower*, and shortly afterwards the word *rose;* for a long time the concept "flower," though more widely applicable than "rose," cannot be said to be more general for the child. It does not include and subordinate "rose"—the two are interchangeable and juxtaposed. When "flower" becomes generalized, the relationship of "flower" and "rose," as well as of "flower" and other subordinate concepts, also changes in the child's mind. A system is taking shape [p. 92–93].

These observations by Vygotsky, based upon his systematic study of children, supply empirical substance to Quine's (1964) logical concept of "semantic ascent":

> It is the shift from talk of miles to talk of "mile." It is what leads from the material (inhaltlich) mode into the formal mode, to invoke an old terminology of Carnap's. It is the shift from talking in certain terms to talking about them [p. 271].

This concept of semantic ascent fits the hierarchy shown in Table 20: from the perceptible qualities of a specific thing, to class names for such things, to classes of these classes, and, lastly, to the dimensions (variables) used in the construction of classes. Each successive level represents a shift from "talking in certain terms to talking about them."

The position of contrast-coordinates at the top of the hierarchy in Table 20 is especially interesting, since in one sense, these represent dimensional concepts of the highest level of abstraction. As Carroll (1964) observes,

> Suppose we take any two words at random, say *tree* and *stone,* and ask a group of people to indicate in what respects these concepts differ, as indeed they must. Among some of the answers we are likely to get are these: A tree is *alive,* while a stone is *inert;* a tree is relatively *flexible,* a stone is *rigid.* That is to say, the mention of any two concepts evokes a series of perceptual or conceptual dimensions in which they differ [p. 102].

Dimensional concepts are to categorical concepts as amino acids are to proteins; the composition or meaning of a category is given by its relative position on a number of dimensions (e.g., Osgood, Suci, & Tannenbaum, 1957). Piaget's description of the critical role of "reversible operations" in propositional thinking (Inhelder & Piaget, 1958) and Kendler's (1965) studies of the slow development of "reversal shift" in discrimination learning of children also indicate the significance of dimensional constructs for sophisticated thinking. Such cognitive developments may not directly influence responses in a word-association experiment, but some parallels may be noted. For example, Entwistle, Forsyth, and Muuss (1964) found that kindergarten children gave predicate and contrast-coordinate associates to adjectives in the ratio, 438 to 104; first-grade children gave about an equal number, 291 to 257; in the third grade, the ratio heavily favored the contrast-coordinates, 102 to 492; and even more so in the fifth grade, 88 to 524.

Whether individual differences in "preference" for one or another of these word-matching bases in the word-association experiment reflects other more general individual differences in cognitive structure has yet to be determined, of course. A generalized difference in "attitude toward words" has been advanced to account for the object-referent and concept-referent sets: object-referent set subjects were hypothesized to treat words generally as an aggregation of labels for concrete things; concept-referent set subjects were hypothesized to treat words generally as concepts within a systematic network of other concepts (Moran et al., 1964). Schmidt (1965) recently tested this rationale by administering a list of homonyms to subjects who previously had been shown to manifest one of the above two sets, on a different word list. As predicted, object-referent set sub-

jects more frequently interpreted the homonyms as labels for things, for example, NUN, sister, and the concept-referent set subjects more frequently as concepts, for example, NONE, nothing. That linguistic habits such as these might possibly have quite broad implications for the investigation of cognitive structure is suggested by studies of semantic conditioning. Riess (1946) conditioned an "electro-dermal response" to selected words in a list, using subjects at four different age levels. Generalization of the conditioned response to homophones, antonyms, and synonyms was then measured. He found, for example, more generalization to homophones than to synonyms (159 versus 129) in the 7.9 age group, but more generalization to synonyms than to homophones (148 versus 52) in the 18.6 age group. Riess (1946) concluded that,

> The present experiment has demonstrated that the relative strength of the semantic gradients does not depend on any a priori quality of the stimulus, but upon the way in which the whole organism utilizes language in its development [p. 151].

Word-Association Commonality and Idiodynamic Sets

One final comment on a methodological issue. The term "commonality" often is used in two quite distinct ways. First, the commonality score of an individual subject has been used to measure the degree to which that individual's associates conform to those given by "normal" subjects. Second, the relative frequencies of stimulus-response word pairs given by a normative group has been used as an index of the relative "association strength" (commonality value) of the word pairs. Idiodynamic associative sets have a marked influence upon both of these commonality measures.

The commonality score of an individual subject, as indicated by several studies in this report, is a partial function of the individual's idiodynamic set and the set compatibility of the stimulus word list. The arbitrariness of commonality scores calculated without regard to this function may account in part for past failures to relate commonality scores to other cognitive or more general personality measures (e.g., Block, 1960; Jenkins, 1960).

Use of normative group frequencies of stimulus-response word pairs as a measure of the relative association strengths of the word pairs (commonality tables) implies a universality of associative tendencies among people of the same linguistic community (Rubenstein & Aborn, 1960). There does seem to be a remarkable consensus in associates to certain stimulus words, like TABLE, chair; BLOSSOM, flower (Russell & Jenkins, 1954). Even though the Kent-Rosanoff word list happens to contain a large number of such words, they are actually extremely rare in the English language (Johnson, 1956). Most stimulus words evoke a wide variety of low-frequency associates. Since the frequency with which a word occurs as an associate (Kent-Rosanoff frequency) and the probability of its emission in general linguistic usage (Thorndike-Lorge count) correlate .94 (Howes, 1957), it follows that the vast majority of words would evoke no one particularly "popular" associate from a normative group. Very likely, then, the commonality hierarchies of most words would resemble those shown in Table 4: that is, largely determined by the proportions of set "types" in the normative sample. While the averaged frequencies shown in such tables may be of use in linguistics, they can be misleading as psycholinguistic indexes of "association strength" of stimulus-response word pairs for persons who use the language. This conclusion only reinforces an earlier one reached by Schwartz and Rouse (1961):

> Applying Whorf's thesis to our results, we could say that word-association responses represent "associations," which, when measured across sufficiently large numbers of people, tend in the aggregate to constitute a measure of linguistic "connections." . . . At present, we have to conclude that the use of group norms to study thought processes is a risky procedure [pp. 99–100].

Summary

An initial formulation of idiodynamic associative sets depicted three groups of subjects who entered the word-association experiment with a definite tendency to give predominately one class of associate, regardless of the stimulus words used. One group of such subjects tended to give syno-

nym-superordinate, another to give contrast-coordinate, and the third to give "functional" (e.g., FOOT, shoe) associates.

In the present study, two 40-word lists were constructed to be equally compatible with these three sets, that is, lists which evoked associates of the above three types with equal frequency from a normative group. When applied to a sample of 482 freshmen students, the lists were found to be reasonably equated for the three sets. A factor analysis revealed the same three factors found in earlier studies. Also, as predicted, subjects with specific sets were shown to achieve their highest commonality scores on set-compatible stimulus words. In collaboration with Dr. Rafael Núñez, the same associative sets and their differential effects upon commonality scores were demonstrated in Spanish-speaking students at the National University of Mexico.

A fourth idiodynamic set, Jung's "predicate type," was then investigated in a 327 freshman sample. The same three factor structure was found for this sample, with the predication variable (adjective-noun, noun-adjective associates) loading highest at the opposite pole of the contrast-coordinate factor. This study was replicated with a sample of 353 freshmen.

In a series of studies, reliability of factor structure, factor scores, and the individual variables representing the four sets was determined. Both internal consistency and test-retest reliabilities generally were adequate. Sex differences were found not to affect factor structure.

Following Siipola's work on task attitudes, the effect of time-pressure instructions upon sets was examined. Lists A and B were administered to two samples of 240 each (with lists reversed in one sample) under time pressure (4-second intervals) and under relaxed conditions (8-second intervals). Siipola's findings of more contrast and fewer predication, with lessened imagery, under time pressure, were confirmed. Time pressure had the following effects upon all sets: (*a*) fewer set-representative associates, (*b*) more associates of the other set-types, (*c*) increase in commonality, (*d*) decrease in predication associates, and (*e*) decrease in reported imagery. These findings led to a redefinition of the so-called speed set (contrast-coordinate) as a dimension-referent set. The perceptual-referent (predicate) set was little affected by time pressure; the dimension-referent set was "weakened" somewhat; but the object-referent (functional) and concept-referent (synonym-superordinate) sets virtually disappeared under time pressure.

The four sets were discussed as generally used bases for matching words, similar to the bases used in object-sorting tasks; and a hierarchy of such word-matching bases, in order of increasing "linguistic sophistication," was proposed.

Concluding remarks concerned the arbitrariness of word-association commonality norms as a measure of "association strength" of stimulus-response word pairs.

APPENDIX

The 40 stimulus words in each of the two lists used in the present studies are given below. They have served a useful purpose in these studies, but at the same time, they have been found not as closely equated as might be desired for future studies.

List A: MILK, SALT, SHIP, JOY, BREAD, SHOVEL, WHISTLE, HAM, SICKNESS, STEEL, BEAUTIFUL, DOCTOR, PLAIN, FIDDLE, DIE, JAM, MASK, ROUGH, PEER, SMILE, GREEN, PAIN, BOY, BLOSSOM, LAMP, CALF, CROW, TABLE, MAN, SHACK, SCISSORS, LONG, EAGLE, SCAB, SOUR, TUG, TOBACCO, HIGH, STRANGLE, KNIFE.

List B: WHISKEY, BUTTERFLY, STOOL, GLARE, STOMACH, JUSTICE, AIL, DISCHARGE, HEAT, OIL, MUTTON, DOCK, INCREASE, TABLET, FOOT, PAIL, HIT, RADIO, YELLOW, END, CAST, STREET, BITTER, STUD, BARK, SOFT, RIP, NAIL, BLACK, CRATE, THIRSTY, WOMAN, COTTAGE, SCALE, SWEET, FRIGID, NEEDLE, SHORT, AFRAID, EATING.

REFERENCES

BLOCK, J. Commonality in word association and personality. *Psychological Reports,* 1960, 7, 332.

BRUNER, J. S., & OLVER, R. R. Development of equivalent transformations in children. In R. C. Anderson & D. P. Ausuble (Eds.) *Readings in the psychology of cognition.* New York: Holt, Rinehart, & Winston, 1965. Pp. 415-434.

CARROLL, J. B. *Language and thought.* Englewood, N.J.: Prentice-Hall, 1964.

DUNN, S., BLISS, J., & SIIPOLA, E. Effects of im-

pulsivity, introversion, and individual values upon association under free conditions. *Journal of Personality*, 1958, **26**, 61–76.

Entwistle, D. R., Forsyth, D. F., & Muuss, R. The syntactic-paradigmatic shift in children's word associations. *Journal of Verbal Learning and Verbal Behavior*, 1964, **3**, 19–29.

Esper, E. A. A contribution to the experimental study of analogy. *Psychological Review*, 1918, **25**, 468–487.

Horton, D. L., Marlowe, D., & Crowne, D. P. The effect of instructional set and need for social approval on commonality of word association responses. *Journal of Abnormal and Social Psychology*, 1963, **66**, 67–72.

Howes, D. On the relation between the probability of a word as an association and in general linguistic usage. *Journal of Abnormal and Social Psychology*, 1957, **54**, 75–85.

Inhelder, B., & Piaget, J. *The growth of logical thinking from childhood to adolescence*. New York: Basic Books, 1958.

Jenkins, J. J. Commonality of association as an indicator of more general patterns of verbal behavior. In T. A. Sebeok (Ed.), *Style in language*. New York: Wiley, 1960. Pp. 307–329.

Johnson, D. M. Word association and word frequency. *American Journal of Psychology*, 1956, **69**, 125–127.

Jung, C. G. *Studies in word association*. New York: Moffat, 1919.

Kelley, T. L. The association experiment: Individual differences and correlations. *Psychological Review*, 1913, **20**, 479–504.

Kendler, T. S. Development of mediating responses in children. In R. C. Anderson & D. P. Ausuble (Eds.), *Readings in the psychology of cognition*. New York: Holt, Rinehart, & Winston, 1965. Pp. 501–520.

Kreezer, G., & Dallenbach, K. M. Learning the relation of opposition. *American Journal of Psychology*, 1929, **41**, 432–441.

Laffal, J. Response faults in word association as a function of response entropy. *Journal of Abnormal and Social Psychology*, 1955, **50**, 265–270.

McGaughran, L. S., & Moran, L. J. "Conceptual level" vs. "conceptual area" analysis of object-sorting behavior of schizophrenic and nonpsychiatric groups. *Journal of Abnormal and Social Psychology*, 1956, **52**, 43–50.

Miron, M. S., & Wolfe, S. A cross-linguistic analysis of the response distributions of restricted word associations. *Journal of Verbal Learning and Verbal Behavior*, 1964, **3**, 376–384.

Moran, L. J., Mefferd, R. B., & Kimble, J. P. Idiodynamic sets in word association. *Psychological Monographs*, 1964, **78** (2, Whole No. 579).

Murphy, G. An experimental study of literary vs. scientific types. *American Journal of Psychology*, 1917, **28**, 238–262.

Osgood, C. E., Suci, G. J., & Tannenbaum, P. H. *The measurement of meaning*. Urbana, Ill.: University of Illinois Press, 1957.

Quine, W. V. O. *Word and object*. Cambridge, Mass.: M.I.T. Press, 1964.

Rabinowitz, R. Learning the relation of opposition as related to scores on the Wechsler Intelligence Scale for Children. *Journal of Genetic Psychology*, 1956, **88**, 25–30.

Reichard, S., Schneider, M., & Rapaport, D. The development of concept formation in children. *American Journal of Orthopsychiatry*, 1944, **14**, 156–161.

Riess, B. F. Genetic changes in semantic conditioning. *Journal of Experimental Psychology*, 1946, **36**, 143–152.

Rosenzweig, M. R. Word associations of French workmen: Comparisons with associations of French students and American workmen and students. *Journal of Verbal Learning and Verbal Behavior*, 1964, **3**, 57–69.

Rubenstein, H., & Aborn, M. Psycholinguistics. *Annual Review of Psychology*, 1960, **2**, 291–322.

Russell, W. A., & Jenkins, J. J. The complete Minnesota norms for responses to 100 words from the Kent-Rosanoff Word Association Test. Technical Report No. 11, 1954, University of Minnesota, Contract N8 onr 66216, Office of Naval Research.

Schmidt, L. Idiodynamic sets and cognitive styles. Unpublished doctoral dissertation, University of Texas, 1965.

Schwartz, F., & Rouse, R. O. The activation and recovery of associations. *Psychological Issues*, 1961, **3** (Whole No. 9).

Siipola, E., Walker, W. N., & Kolb, D. Task attitudes in word association, projective and nonprojective. *Journal of Personality*, 1955, **23**, 441–459.

Tendler, A. D. Associative tendencies in psychoneurotics. *Psychological Clinic*, 1933, **22**, 108–116.

Thurstone, L. L. *Multiple-factor analysis*. Chicago, Ill.: University of Chicago Press, 1957.

Vygotsky, L. S. *Thought and language*. Cambridge, Mass.: M.I.T. Press, 1962.

Watt, H. J. Experimental contribution to a theory of thinking. In J. M. Mandler & G. Mandler (Eds.), *Thinking: From association to Gestalt*. New York: Wiley, 1964, Pp. 189–200.

Wells, F. L. The question of association types. *Psychological Review*, 1912, **19**, 253–270.

Woodworth, R. S. *Experimental psychology*. New York: Holt, 1948.

(Received August 10, 1965)

Vol. 80, No. 5 Whole No. 613, 1966

Psychological Monographs: General and Applied

SIMILARITY RELATIONS AMONG CERTAIN ENGLISH SENTENCE CONSTRUCTIONS[1]

CHARLES CLIFTON, JR. AND PENELOPE ODOM

Institute of Child Behavior and Development, University of Iowa

Vanderbilt University

The implications of generative grammars of English for grammatical relationships among sentence constructions were discussed. A series of studies investigating the psychological similarity of sentence constructions was carried out, employing judgment and recognition techniques. Metric and nonmetric multidimensional scaling methods were used to analyze the data and to analyze data reported by Mehler (1963). A highly consistent pattern of psychological similarity was found among the constructions investigated. The obtained pattern was congruent with the pattern of grammatical relationships that the Katz and Postal (1964) approach appeared to imply.

Recent revolutionary changes in descriptive linguistics, stemming largely from the work of Noam Chomsky in the mid-'50s (e.g., Chomsky, 1957), promise to furnish descriptions of language that are more adequate for the psychologist working in the field of language structure than were the previously available descriptions. The new formulations are also seen by some as furnishing many hints about the workings of a language user. The research to be reported here is concerned with one of the topics treated by the modern generative grammars, that of the syntactic relationships among sentences. Grammarians of Chomskian persuasion claim that their grammars, in contrast to the traditional grammars, clarify the underlying structural relationships among certain sentence constructions that are seen by the language user as intimately related or highly similar. The purpose of the present research is to obtain precise behavioral measures of the similarity of such syntactically related sentence constructions. As careful measures of the perceived similarity of intuitively and grammatically related sentences, the obtained data should have intrinsic interest. Further, the data are relevant to the broad topic of the parallels between the linguistic description of a language and the abilities of the person who uses that language, that is, the relation between grammar and "cognitive structure." Finally, the data may be of value to the linguist who is willing to consider such empirical evidence in evaluating alternative statements of grammatical relations.

The nature of the new grammars, generative grammars, will be discussed in sufficient detail to make clear the types of sentence relationships they posit. No attempt will be made to consider the psychological implications of generative grammars; for indications of such implications, the reader is referred to Miller and Chomsky (1963), Chomsky (1961), and Katz and Postal (1964). The bulk of the mono-

[1] Portions of the material in this report were presented in PhD dissertations submitted by the authors to the University of Minnesota. The authors wish to express their deep indebtedness to James J. Jenkins for the invaluable aid and encouragement he supplied.

We also express our gratitude to J. B. Kruskal for supplying MDSCAL, the scaling program used to analyze much of the data reported in this monograph, and to Donald Boyd for adapting the program to the University of Iowa IBM 7044. The support of the University of Iowa Computer Center in the machine analysis of the data is gratefully acknowledged. Thanks are due also to John Tyler for conducting the machine computation involved in the factor analyses. Part of the present research was supported by Grant G-18690 from the National Science Foundation to James J. Jenkins, University of Minnesota. The factor-analysis computation was supported by Grant RD 846p from the Office of Vocational Rehabilitation to Jum Nunnally and Richard Blanton, Vanderbilt University.

graph will consist of a presentation of the measured similarity of the sentence constructions investigated, primarily in terms of the scaled psychological distances among the constructions.

GENERATIVE GRAMMAR

Phrase Structure and Transformation Rules

The aim of a modern grammar of a language is to provide an explicit enumeration and description of the sentences in the language, relying on purely formal (especially, nonsemantic) statements. A generative grammar accomplishes this aim through the use of rules that can be applied in such a way that they can generate all the sentences in the language, and only those sentences. The generative grammar is to contain a basic axiom and a set of rules specifying that the axiom is to be repeatedly rewritten in certain ways, producing series of strings of symbols. The final string in each series of strings is to be a sentence in the language, and the nonfinal strings in the series may be said to underlie the final sentence. It should be made clear that the rules to be considered here act upon, and produce, symbols and strings of symbols which *underlie* sentences in the language. They do not themselves act upon or produce actual sentences. That is, the grammatical rules generate the structures underlying sentences, not the sentences themselves.

The following unordered set of rules, similar to one presented by Miller (1962), is illustrative of one type of grammatical rule. In this example, → is to be read as "is to be rewritten as" and S is the basic axiom. Italicized multiple terms on the right end of the rewrite arrows, separated by commas, are lists of optional choices.

Given: S

$S \rightarrow NP + VP$

$NP \rightarrow T + N$

$VP \rightarrow V + NP$

NP → *Bill, John, . . .*

T → *the, a*

N → *boy, girl, ball, . . .*

V → *hit, struck, . . .*

With this set of rules, such simple sentences as "The boy struck the girl" and "John hit the ball" can be generated. As mentioned earlier, the grammar is to provide a description of the sentences it enumerates. Properly phrased, the preceding set of rules would assign a structural description to each of the sentences it generates, that is, would indicate which parts of each sentence form functional units or "constituents" and would assign some class label to each constituent. For the first sentence generated in the present example, the rules would indicate that "the boy" and "struck the girl" form units (to be labeled NP and VP, respectively); that "struck" and "the girl" form units (V and NP, respectively); and that "the," "boy," and "girl" form units (T, N, and N, respectively).

The rules of the example represent only one of a number of different types of rules. Rules of this type are termed "phrase-structure" rules, and that part of a generative grammar consisting of them, the "phrase-structure component." Phrase-structure rules have the defining characteristic that they specify that a *single* symbol is to be rewritten as one of a specified set of symbols and that the symbol is to be so rewritten *whenever* it occurs (or whenever it occurs in a specified immediate context). That is, such rules rewrite only single symbols, not strings of symbols. Further, their application is not contingent upon the structure of the string of symbols in which the symbol to be rewritten is embedded.[2]

Another type of rule proposed by Chomsky, the "transformation rule," differs from phrase-structure rules on both these

[2] It should be pointed out again that the phrase-structure rules in a grammar of a natural language, as well as the transformation rules to be considered next, would not actually generate sentences in the standard writing system of the language, as the rules of our example do. Rather, they would generate strings of more or less abstract symbols, representing morphemes or classes of morphemes, which later-applied rules would convert into a phonetic notation. That is, they generate strings which underlie a sentence.

counts. A transformation rule allows *an entire string* (or set of strings) of symbols to be rewritten as a different string. Further, the applicability of a transformation rule to a particular string is contingent upon the structural description of the string; a particular transformation can be applied only to strings of certain specified structural descriptions. An example of a transformation rule proposed by Chomsky (1957) is the passive transformation. The passive transformation specifies that a string of symbols (morphemes or names of classes of morphemes) having the structural description $NP_1 - Aux - V - NP_2$ can be rewritten as $NP_2 - Aux - V - NP_2$ $en - V - by + NP_1$. The transformation can be used to rewrite the string of symbols that underlies, for instance, the sentence "John hit the ball" into the string that underlies "The ball was hit by John." The transformation rule, like the phrase-structure rule, is to be phrased in such a way that it provides a structural description of its output.

A transformation, such as the passive transformation, that specifies how a single string is to be rewritten is termed a "singulary" transformation. Another type of transformation rule, the "generalized" transformation, rewrites a set of two or more strings as a single string. For instance, there is a transformation rule which converts the strings underlying "John hit the ball" and "The ball is green" into the single string underlying "John hit the green ball."

Transformation rules can also be classified as "optional" or "obligatory." An optional transformation, such as the transformation already mentioned, need not be applied whenever a string to which it can be applied is available to it. An obligatory rule, however, like a phrase-structure rule, must be applied whenever it can be applied. An example of an obligatory transformation is one which transforms the string underlying the nonsentence* "They brought in him" (analogous to "They brought in the criminal") into the string underlying "They brought him in."

Generative grammarians agree that phrase-structure rules and obligatory transformation rules are involved in the generation of any sentence, no matter how simple its structure. In the type of grammar thus far described, some optional transformations are used to generate sentences of relatively complex structure. It may be pointed out that the use of optional transformation rules and generalized transformation rules is currently being called into question by some linguists. It is argued that the strings underlying sentences may best be generated by the phrase-structure and the obligatory singulary transformation rules alone (cf. Chomsky, 1965).

Other types of rules are used to convert the products of the phrase-structure rules and the transformation rules, which are the strings of symbols underlying sentences, into sentences in phonetic notation. However, the investigation to be reported here is concerned with syntactic structure, and the essential description of the syntactic structure of a language is provided by the phrase-structure and the transformation rules. For this reason, rules of lower level than these will not be considered here.

Chomsky's Transformational Grammar

Chomsky, in his earlier published works (e.g., Chomsky, 1957, 1961, 1962; Chomsky & Miller, 1963; for an introductory treatment differing in some respects, see Bach, 1964) has provided portions of a grammar of English in which certain intuitively related sentences are described as being transformationally related to each other. In his grammar, the phrase-structure rules apply to the initial symbol and to the products of rewriting the initial symbol. Transformation rules apply to the products of the phrase-structure rules and to the products of other transformation rules. The phrase-structure rules plus the applicable obligatory transformation rules are used to generate the strings underlying some of the simplest, most common, sentences in the language. These strings are called "terminal strings" and the corresponding sentences, "kernel sentences." In English, these are the simple, active, declarative sentences such as "John hit the ball."

Optional transformation rules are applied to the terminal strings (not to kernel sentences) and are used to produce the more complex sentences in the language. Examples of optional singulary transformations are the passive transformation, which produces strings underlying passive sentences such as "The ball was hit by John"; the negative transformation, which allows the derivation of negative sentences such as "John didn't hit the ball"; and the question transformation, which produces yes-no questions such as "Did John hit the ball?" In some cases, transformation rules can be applied to the products of other transformations. For instance, the negative transformation can be applied to the product of the passive transformation (yielding, eventually, the passive-negative sentence, "The ball wasn't hit by John"), and the question transformation can be applied to the products of either or both the passive and the negative transformations (producing the passive question, "Was the ball hit by John?"; the negative question, "Didn't John hit the ball?"; and the passive-negative question, "Wasn't the ball hit by John?").

A kernel sentence and its corresponding passive represent a pair of sentences that Chomsky sees as intuitively related, intuitively similar structurally. In a grammar such as Chomsky's, they are also seen to be transformationally related. They have essentially identical histories of derivation, except that the passive transformation was applied to only one of the pair of sentences. The application of the passive transformation in the derivation of the kernel would make it a passive. Similar transformational relationships can be seen to exist between the kernel and the negative, the kernel and the question, the passive and the negative question, etc. One can speak of a sentence family, which is defined by a particular set of optional singulary transformations and a single terminal string, and which consists of all the sentences which result from the application of all permissible combinations of the set of transformations and no other optional transformations. The absence of any of the specified optional transformations will be considered to be a permissible combination of transformations; thus, the kernel is a member of each sentence family.

TABLE 1

EXAMPLE OF P,N,Q SENTENCE FAMILY

Sentence	Construction
The man closed the box.	K
The box was closed by the man.	P
The man didn't close the box.	N
Did the man close the box?	Q
The box wasn't closed by the man.	PN
Was the box closed by the man?	PQ
Didn't the man close the box?	NQ
Wasn't the box closed by the man?	PNQ

Table 1 displays the eight members of the sentence family defined by the passive, the negative, and the question transformations, and the terminal string underlying "The man closed the box." It is to be noted that the eight sentences used in a previous paragraph as an example of the result of applying the passive, the negative, and the question transformations form a similar sentence family, with "John hit the ball" as kernel. For brevity, such sentence families will be referred to as "P,N,Q" families, and the grammatical constructions of their member sentences will be abbreviated as K (kernel), P (passive), N (negative), Q (question), . . . , PNQ (passive-negative question).

The grammatical relationships among the members of a sentence family can be described by comparing their transformational histories. Those sentences that differ by a single transformation would be the most closely related, while, in the present example, those differing by all three transformations would be the least closely related. Miller (1962) has furnished a graphic way of presenting the transformational relationships among sentence family members. The members of a P,N,Q sentence family can be represented in the cube of Figure 1. The closeness of the relationship between two sentences (two vertexes in the diagram) is indicative of their degree of syntactic relationship, and is measured as a function of the number of lines in the

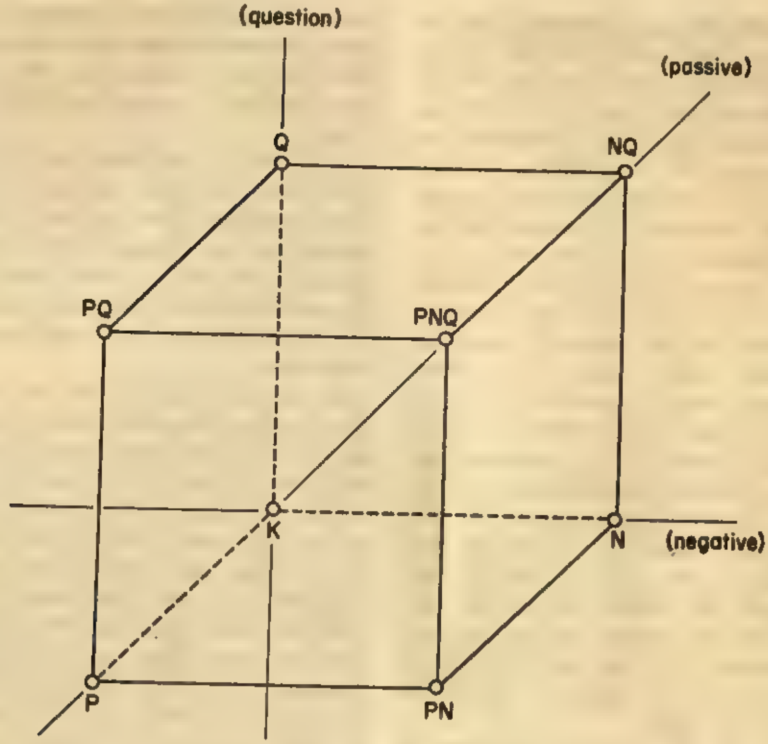

Fig. 1. Cube representing transformational relationships among sentences.

shortest path connecting them. Note that this cube is non-Euclidean; specifically, diagonals in the figure are undefined. In traveling between two nonadjacent vertexes of the cube, one must move through some set of the intervening vertexes. This property of the cube reflects the step-wise property of the transformational grammar. That is, in moving from one sentence construction to a construction removed by, say, three transformations from the first, one must pass through strings underlying constructions removed by one and by two transformations. If it is assumed that the distance between any two vertexes is equal to the sum of the lengths of the shortest series of lines connecting them and that all corners in the figure are right angles, the distances in the configuration are equal to distances in a "city block" space (Attneave, 1950). That is, the distance between any two points is the sum of the absolute values of the differences between their projections on each dimension. In Miller's cube, one transformation would be coordinated with each dimension.

An Alternative Generative Grammar

There is some doubt among linguists that the K, the N, the P, etc., sentences are best viewed as transformational variants of the same sentence or the same underlying terminal string. Rather, some or all of the sentences, it is argued, should have different derivations in the phrase-structure component of the grammar.

For instance, Lees (1960) writes his phrase-structure rules in such a fashion that they, rather than the transformation rules, optionally introduce a negative morpheme (as a slightly aberrant member of a class of "preverbs") in the generation of a sentence. If the negative morpheme is produced in the phrase-structure portion of the derivation of a sentence, then the sentence will be a negative; if not, then the sentence will be nonnegative (although preverbs other than "not," such as

"never," or "seldom," or "always," may be chosen, even if "not" is not chosen). There still is a negative transformation, but it acts only on sentences having underlying strings containing the "not" morpheme (or some other preverb) and simply changes the order of the morphemes in the string. In such cases, the negative transformation is now an obligatory transformation.

Katz and Postal (1964) provide similar treatments for the passive and question sentences, as well as for negative sentences. Phrase-structure rules, rather than transformation rules, are to introduce a passive morpheme or a question morpheme in the derivation of a passive or a question sentence. Again, there will presumably still be passive and question transformations, but, as was the case with the altered negative transformation, they will be obligatory, applicable only to strings having the appropriate phrase-structure derivation, and will serve primarily to change the order of the elements of the string and to delete certain of the elements.

The treatment of negative sentences as having different phrase-structure derivations than nonnegative sentences is generally viewed as being preferable to the transformational treatment of negatives. However, the analogous treatment of the remaining sentence constructions is more open to question. In fact, although they provide some grammatical justification for their treatment, Katz and Postal (1964) provide different phrase-structure derivations for passives, questions, etc., for primarily semantic reasons. They wish to provide a description of both the syntactic and semantic aspects of a language, using a transformational grammar for the former and Katz and Fodor's (1963) type of semantic theory for the latter. Loosely speaking, such a semantic theory provides rules used to interpret sentences semantically, and Katz and Postal (1964) argue that (in the case of sentences having no generalized transformations in their derivation) these interpretative rules should apply to the structured strings produced by *only* the phrase-structure rules. Since the kernel and the question sentences, say, obviously have different meanings, they must have different underlying strings on the phrase-structure level and therefore must have different phrase-structure derivations.

It is not appropriate here to evaluate Katz and Postal's approach. However, their approach does contain implications concerning the relationships among sentences which differ from those of other grammars. This section will be devoted to an examination and interpretation of these implications.

An obvious attack on the problem of determining the relationships among sentences within the Katz and Postal framework is to claim that sentences are related, or similar, as some function of the extent to which their underlying structured strings are similar. Specifically, the relationship among sentences would be a function of the extent to which their underlying strings have many morphemes in common. This approach indicates that "John hit the ball" and "The ball was hit by John" would be about as closely related to one another as "John hit the ball" and "The ball hit John" are to each other, since the sentences of each of these pairs share many of their morphemes. However, the linguist wants to argue that the sentences of the former pair are linguistically related versions of each other, while the sentences of the latter pair are not.

Katz and Postal (1964) propose one resolution to the problem of distinguishing between those sentences that are grammatically related and those that are not. They indicate that sentences whose underlying strings differ only in morphemes that are linguistically universal are the sentences that are closely related grammatically. Linguistically universal morphemes are morphemes that are specified by a general theory of linguistic descriptions, rather than being morphemes specific to certain languages. The negative morpheme, the question morpheme, and the passive morpheme, they claim, are likely to be such universal morphemes. Thus, it might be that the active and the passive, the question and the nonquestion, etc., each differ

from one another in only one universal morpheme, and thus that the grammatical relationships among them are much the same as the relations that existed when they were analyzed as transformational variants of the same terminal string.

The solution, unfortunately, does not appear to be this simple. For instance, in the Katz and Postal (1964) approach, the question and the kernel are viewed as differing by more than the single universal question morpheme. They are analyzed as differing also in the possession, by some strings involved in the derivation of the question sentence, of "wh" (another universal morpheme) and "either-or" (a morpheme peculiar to English). It is suggested (Katz & Postal, 1964, p. 119) that a question like "Is John a doctor?" results from the disjunction between the string underlying "John is a doctor" and the string underlying "John is not a doctor." In fact, the question sentence is formally related to the disjunctive structure underlying "Either John is a doctor or not." One of the components of this disjunction, it is argued, is a structured string underlying the kernel, and the other a structured string underlying the negative. Any relationship between the question and the kernel, or the question and the negative, is to be accounted for on the basis of this part-whole relationship. Thus the question is grammatically related both to a string underlying the kernel and to a string underlying the negative. It appears that the question would be more or less intermediate in closeness of grammatical relationship to the kernel and the negative.

This interpretation of the relationship between questions and nonquestions must be considered tentative, and a similar reservation must be made regarding the interpretations of the relationships among the other sentence structures in the Katz-Postal approach. Probably affirmative and negative sentences are to be considered as differing in only a single universal (negative) morpheme, and thus as directly related. A similar analysis apparently may be made of the active and the passive. It may be that the passive morpheme and the negative morpheme can occur together in a combinatorial fashion and that a kernel and a passive-negative sentence, or a passive and a negative sentence, differ by just two universal morphemes in their underlying strings. Such a treatment would indicate relations among the K, the N, the P, and the PN sentences (the nonquestions) that are very similar to the relations indicated by the transformational treatment.

This is not the case, however, for the relations among the questions. It appears as if a negative morpheme and a question morpheme are to occur both in the string underlying an affirmative question and in the string underlying a negative question. Perhaps "did" and "didn't," and "was" and "wasn't," in affirmative questions and negative questions, are simply alternative ways of writing the same element in an underlying string. Thus, there would be very little if any structural difference between affirmative questions and negative questions, such questions perhaps differing only stylistically. Passive questions and active questions, on the other hand, seem to differ in the possession of a universal passive morpheme, and thus to be related to each other in the same fashion as passive and active nonquestions.

It may be legitimate to represent, at least tentatively, the sentence relationships posited by the nontransformational approach as the prism of Figure 2, analogous to the cube of Figure 1. This prism differs from the cube primarily in the placement of the affirmative and the negative questions close to one another, and midway between the affirmative and the negative nonquestions. On the assumption that questions that differ by the possession of the passive morpheme (Q and PQ; NQ and PNQ) are neither more nor less closely related than nonquestions so differing, the prism is given parallel active and passive planes. It is not perfectly clear that the figure should be three-dimensional, with the questions removed from the nonquestions; it might be legitimate to view the questions as falling in the same plane as the nonquestions and thus to represent all eight sentence constructions in a two-dimensional figure.

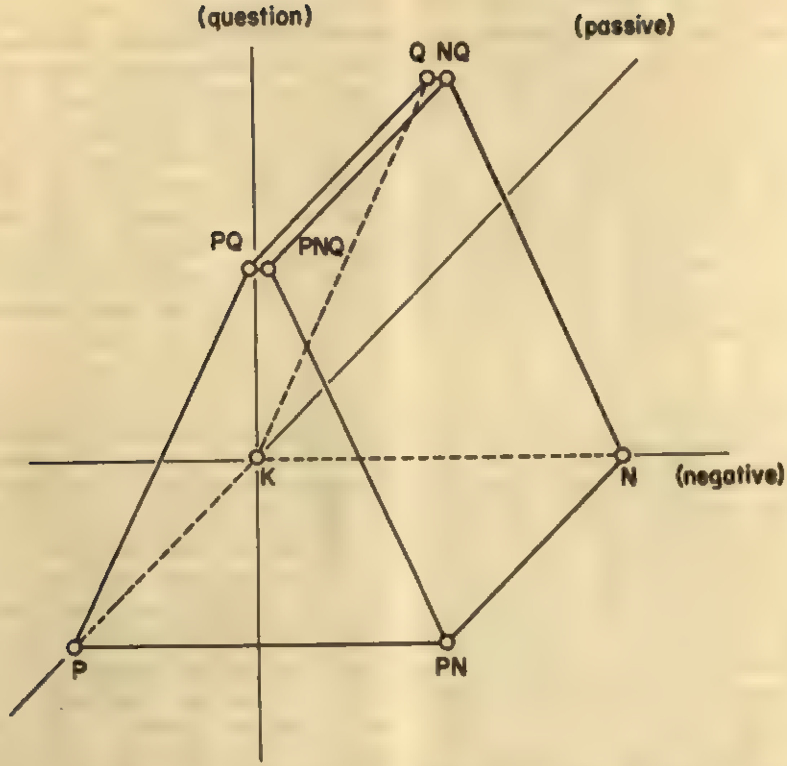

FIG. 2. Prism representing phrase-structure relationships among sentences.

The cube representing the transformational relationships among sentences was said to be non-Euclidean, reflecting the step-wise (i.e., number of transformational steps) nature of the relationships. On the other hand, the Katz-Postal phrase-structure relationships are essentially "common elements" (or "differences in elements") relationships. Such relationships do not so unequivocally suggest a step-wise, non-Euclidean representation. It would seem most reasonable to view the prism representing the phrase-structure relationships among the sentences as a Euclidean solid, at least in the absence of closer analysis of the relationships.

STUDIES OF SENTENCE SIMILARITY

Two alternative representations of the grammatical relatedness of sentence constructions have been presented. It must be realized that these representations, especially the syntactical relationships indicated by the nontransformational, phrase-structure approach, are to be regarded as tentative interpretations of incompletely worked-out grammars. Further, these representations have been considered to be metricized, perhaps to an illegitimate degree; there may be no justification for speaking of the relationships as distances, Euclidean or non-Euclidean. Finally, the hidden assumption used to construct the representations, namely that the unit of distance coordinated with a transformation (or universal morpheme) is the same, regardless of which transformation (or morpheme) is involved, has no justification in the grammars. In fact, one might be willing to argue that the grammars indicate that certain distances should be greater than others. For instance, the transformational approach analyzes the passive transformation as being more complex than the negative transformation, in the sense that more of the elements of underlying strings are changed by the former than by the latter. If transformational complexity is to

be coordinated with distance, then the distance between actives and passives should be greater than the distance between affirmatives and negatives. However, the grammars do not speak clearly enough on the topic of the closeness of the relationships between pairs of constructions differing by a single transformation to allow any sure predictions to be made regarding the relative distances coordinated with the various transformations or differences in phrase-structure derivation.

Given these reservations, one can ask if various behavioral measures of the similarity of sentences map out one of the representations of the grammatical relationships among the sentences. To provide an answer, a series of studies investigating the similarity among sentences was carried out. The sets of sentences investigated were those which the transformational approach analyzes as members of P,N,Q sentence families. (For ease of discussion, the vocabulary of a transformational grammar will be used throughout to describe the sentence constructions and the relations among them.) The studies measured the judged similarity of such sentences, the confusion between such sentences in a recognition task, and the changes in confusion with repeated trials on the recognition task. Further, some data presented by Mehler (1963) on confusions among such sentences in a recall task were reanalyzed.

Method of Analysis

It seems appropriate to describe here the method of analysis to be used. The basic data from Experiments 1 and 2, and the reanalysis of Mehler's (1963) experiment, can be represented in an 8×8 matrix of the measured similarities or dissimilarities between each of the 64 possible pairs of the eight sentence constructions (K, P, N,..., PNQ). The data from Experiments 3 and 4 can be represented similarly in 4×4 matrices. Such data representations are too complex to be comprehensible, and statistical analyses of the data are similarly undesirably complex.

It is possible to achieve an efficient reduction of such data through multidimensional scaling procedures, thereby determining spatial representations of the sets of sentences, such that each sentence construction corresponds to one point in the spatial configuration and that the distances among the points are related to the measured similarities of the sentences (Shepard, 1960, 1962a, 1963). Such spatial representations are very suitable for testing the implications of the grammatical descriptions, which were themselves given spatial representations.

Metric multidimensional scaling techniques (Torgerson, 1958) were used in some of the studies to determine the configurations appropriate for the sets of obtained similarity measures. In the remaining studies, a nonmetric scaling technique developed by Kruskal (1964a, 1964b) was used. In this technique, the matrices of similarities or dissimilarities serve directly as the input to the scaling program. For any specified number of dimensions in which the data scaling is to take place, the technique determines a set of distances among the points being scaled which satisfy the metric axioms governing physical distance. That is, it determines a configuration of points which can be interpreted as a spatial model of the objects being scaled. It outputs the coordinates of each point in the configuration, as well as the interpoint distances. The configuration determined by the program is such that the interpoint distances in it bear the closest possible montonic relationship to the similarity (or dissimilarity) measures used as input. No assumption about the function relating the similarity measures to the interpoint distances is made, except that the function is a monotonic one. That is, for one distance in the resulting configuration to be greater than another distance, the measured similarity corresponding to the former distance must be less than the similarity corresponding to the latter. The similarities are thus assumed to be measures on only an ordinal scale, while the output distances satisfy the requirements of a ratio scale.

Kruskal's procedure furnishes a measure of how well the distances fit the similarity measures, that is, how closely the function relating distances to similarities approaches monotonicity. This measure of the "success" of the scaling is the normalized residual variance of the monotone regression of distance upon similarity, and is termed *stress*. As the technique is an iterative one, the stress obtained for any set of data in a given dimensionality is to some extent a function of the number of iterations employed. However, by increasing the number of iterations, any desired degree of approximation to the absolute minimum stress can be reached. As a rule of thumb, Kruskal (1964a) indicates that a final stress of 10% is to be considered "fair," 5% "good," 2½% "excellent," and 0% "perfect."

The technique allows scaling to take place in a variety of non-Euclidean spaces, as well as in Euclidean space. Of greatest interest is the possibility of scaling the similarity data in a city block space (Attneave, 1950). Such a capacity is of value, because, as it will be recalled, the distances in the city block spatial model are equal to the "around-the-edges" distances of the transformational relationships representation of Figure 1, when all angles in the solid are right angles.

Experiment 1[3]

Method

The first study took the most obvious approach to determining the similarity among sentences, that of having the subjects (*S*s) simply rank the similarity of the members of sets of K,P,N,..., PNQ sentences. The procedure used was the method of multidimensional rank order (Torgerson, 1958).

Two groups of introductory psychology students at the University of Minnesota were run in a group setting, $N = 55$ in Group 1 and $N = 43$ in Group 2. Each *S* was given a booklet containing 16 sets of eight sentences each. The eight sentences in a set were the eight members of a P,N,Q sentence family; each set of sentences represented a different P,N,Q family. One member of each of these 16 families was set apart as a "Standard" sentence. There were two Standard sentences in each of the eight sentence constructions. The Standard sentences are listed in Table 2. Note that in Experiment 1, unlike the other experiments to be reported, the combination of "do" plus the negative morpheme was not contracted.

The *S*s were instructed to rank the seven remaining members of each set of sentences with respect to their similarity to the Standard sentence of the set, by assigning the number 1 to the sentence most similar to the Standard, the number 2 to the next most similar, etc. The *S*s were instructed to rank in terms of the similarity of the sentences as they understood "similarity." It was emphasized that there were no right or wrong answers. The task was administered as a paper and pencil test, with a liberal 1-hour time limit.

The two groups of *S*s were given the same sets of sentences to rank, in the same order; the two groups were used simply for the purpose of determining the reliability of the rankings. The same random order of presentation of sentences families, and order of presentation of sentence constructions within a sentence family, were used for all *S*s.

The data were analyzed first in terms of the reliability of the mean ranks between the two groups and between the two Standard sentences in each construction. Following this, comparative distances between the pairs of constructions were obtained from the data of Group I only, using the method described by Torgerson (1958). These comparative distances were treated as ordinal measures of the dissimilarity of the sentences and scaled using the Kruskal technique described earlier. Also, under the assumption that they were measures on a linear scale of the true interpoint distances, the comparative distances were converted to distances on a ratio scale by adding a constant, and the configuration of sentences corresponding to these distances was determined.

TABLE 2

STANDARD SENTENCES USED IN EXPERIMENT 1

Betty did not read the book.
His mother prepared the dinner.
Did the horse eat the hay?
Was the pipe dropped by the plomber? (sic)
The tooth was not filled by the dentist.
Did not Mary see the fish?
The shirt was sold by the clerk.
John hit the ball.
Was not the piano moved by the truck?
The paper was written by the student.
The house was not built by the carpenter.
Was not the bell rung by Bob?
Was the cat chased by the dog?
The man did not close the box.
Did the professor give a lecture?
Did not the woman spank the child?

Results

Reliability. Both the reliability over the two groups of *S*s and the reliability over the two examples of Standard sentences in a particular construction were determined. A separate computation of reliability was made for each construction of the Standard sentences (that is, the reliability of the ranks assigned the seven sentences ranked with a K as Standard sentence, the reliability of the ranks assigned with P as Standard, etc., were all calculated). In determining the reliability between the two groups for Standard sentences in a particular construction, the ranks assigned sentences by each group were averaged over the two examples of Standard sentences in that construction and over *S*s in the group, and the averaged ranks for Group 1 were correlated with the averaged ranks for Group 2. Eight such correlations were determined, one for each sentence construction that served as Standard. In calculating the reliability between the two examples of Standard sentences in each construction, the ranks assigned sentences for one of the examples were averaged over all the *S*s in the two groups and correlated with the similarly averaged ranks for the other example. Again, eight correlations were obtained.

[3] Ida Kurcz, of the University of Warsaw, collected and performed the preliminary analyses on the data reported as Experiment 1, during a postdoctoral year at the University of Minnesota. The authors wish to express their thanks to her for making her data available to them.

TABLE 3
COMPARATIVE DISTANCES, EXPERIMENT 1

	K	P	N	Q	PN	PQ	NQ
P	−1.81						
N	1.48	1.64					
Q	−0.28	0.14	0.43				
PN	1.65	1.18	−0.98	0.55			
PQ	0.24	−0.14	0.67	−1.26	0.34		
NQ	0.16	0.22	0.15	−0.92	0.29	−0.88	
PNQ	0.28	−0.03	0.53	−0.80	0.19	−0.99	−1.29

The reliability of the averaged ranks was satisfactory. The correlations between the two groups ranged from +.97 to +.99, as did the correlations between the two examples of Standard sentences.

Scaling. As mentioned earlier, only the data from Group 1 were scaled. Comparative distances between the pairs of sentence constructions were obtained using the method described by Torgerson (1958, pp. 263–269). This method may be considered an extension of the Thurstone comparative judgment technique. Sixteen incomplete 8 × 8 matrices (one for each Standard sentence), containing the frequencies with which *S*s judged one sentence (i) in a family to be more similar to the Standard sentence (k) of the family than is another sentence (j) in the family, were obtained. An *S*'s set of ranks for one sentence family was discarded if he made an error in the ranking, for example, if he assigned the same rank twice. No more than two *S*s had to be discarded for any sentence family. The corresponding frequencies ${}_{k_1}F_{ij}$ and ${}_{k_2}F_{ij}$ in the pairs of matrices based on the two standard sentences of one construction were summed and divided by the total usable N for the pair of matrices, to obtain eight incomplete 8 × 8 matrices of proportions, ${}_kp_{ij}$. These proportions were converted to normal deviates, which were used, following Torgerson's (1958) least-squares procedure, to determine a single symmetric 8 × 8 matrix of comparative distances between the pairs of sentence constructions. These comparative distances are presented in the halfmatrix of Table 3.

The halfmatrix of comparative distances (with the value 2.0 added to all entries, to make them all positive) was scaled using Kruskal's technique. These dissimilarities were found to scale in one dimension with a stress of 0.00% (necessarily, in both Euclidean and city block spaces). K and P occupied identical positions in the configuration, as did N and PN, and the four questions fell on a single point halfway between the affirmative and negative nonquestions. The plot of interpoint distances in the configuration against the input dissimilarities is presented in Figure 3. The plot may be termed a step function. Apparently the data have a property, mentioned by Shepard (1962b), that results in somewhat unsatisfactory solutions. When the points in subsets of the whole set of points are more similar to one another (closer together in the underlying configuration) than they are to any points outside the subset, the scaling technique will indicate that all the points in each subset have the same coordinates in the resulting configuration, that is, are identical. Such a

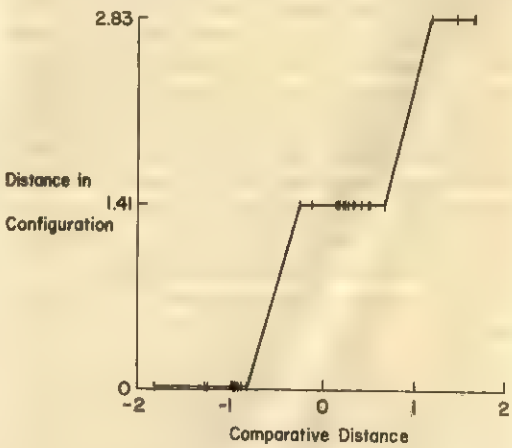

FIG. 3. Distance in configuration versus comparative distance, Experiment 1.

situation is likely to result in an excessively low dimensionality of the configuration. In the case where there are just two such subsets, the data will necessarily scale in one dimension. When there are three such subsets, the data will scale in at most two dimensions. In the present case, there seem to be three such subsets: K,P; N,PN; and Q,PQ,NQ,PNQ. The technique thus failed to discriminate between sentences differing by just the passive transformation, and between all the question sentences.

Some conclusions may be drawn from the obtained configuration, regardless of its shortcomings. The position of the questions midway between the affirmatives and the negatives is congruent with the Katz and Postal (1964) grammatical analysis, as is the lack of any distinction between affirmative questions and negative questions. The disappearance of any distinction between sentences due to the passive-active difference is congruent with neither of the proposed models, although it may be interpretable in terms of the lack of any difference in meaning between passives and actives. However, it is plausible that the passive-active difference in the obtained dissimilarity measures may simply be a good deal smaller than the affirmative-negative difference and that the present divisibility of the sentences into homogeneous subsets has obscured the relatively small passive-active difference.

The Kruskal technique uses only the ordinal information present in the data. The comparative distances used as input data, however, are presumably measures on an interval scale. These comparative distances may be scaled using metric scaling techniques described in Torgerson (1958), as suggested by Shepard (1962b). Such a procedure should allow a more sensitive determination of any passive-active difference present in the data.

The comparative distances of Table 3 must be converted to distances on a ratio scale by adding a constant, in order to determine the multidimensional configuration corresponding to them. The necessary additive constant was estimated by the procedure of Messick and Abelson (1956) to be equal to 2.527. The distances resulting from the addition of this constant to the comparative distances are the distances between the points (sentence constructions) of the configuration of sentence constructions, under the assumptions required by the scaling technique used. The coordinates of the points in this configuration were determined by calculating the matrix of the scalar products of the centroid-origin vectors of the eight points in the configuration, using the procedure suggested by Torgerson (1958, pp. 254–259). This matrix of scalar products was factored using the principal-axes method. The loadings of each point on the first n factors extracted are the coordinates of the points on the largest n dimensions of the configuration.

The factor solution obtained indicated that the configuration of sentences could be considered to exist in a real space. One negative latent root was obtained, but its absolute value was only 3% of the value of the sum of the positive latent roots and could thus be considered to be well within the limits of error.

Further, the solution indicated that the configuration could be fairly adequately represented in three dimensions. Following the method of Torgerson (1958, pp. 278–279), the total variance of a scalar product matrix derived from the first three factors extracted was compared with the total variance of the scalar product matrix which was factored. The sum of squares of the elements of the matrix derived from the first three factors equaled 96% of the value of the sum of squares of the original scalar product matrix.

The three-dimensional configuration of sentences whose coordinates equaled the loadings on the first three factors was translated and rotated to a position such that the coordinates of the K construction were (0,0,0); the coordinates of the P construction (x,0,0), where x indicates simply that the coordinate of P on the first dimension was free to vary; and the coordinates of the N construction (y,z,0). The resulting coordinates are presented in Table 4. The dimensions of the rotated

TABLE 4
ROTATED FACTOR LOADINGS, EXPERIMENT 1 (PROJECTIONS OF POINTS OF SENTENCE CONFIGURATION ON THE DIMENSIONAL AXES)

Construction	Dimension		
	Passive	Negative	Question
K	0.00	0.00	0.00
P	1.26	0.00	0.00
N	0.33	4.06	0.00
Q	0.10	1.67	1.52
PN	1.87	3.73	0.06
PQ	1.19	1.62	1.74
NQ	0.53	2.03	1.58
PNQ	1.32	1.79	1.64

configuration may be identified as a passive dimension, a negative dimension, and a question dimension. The projections of the configuration on each pair of dimensions are displayed in Figure 4, together with an oblique projection of the entire configuration (with the passive dimension foreshortened by $\sqrt{2}$ for clarity of presentation). These projections may be compared with the predicted configurations, Figures 1 and 2.

There seems to be substantial similarity between the obtained configuration and that predicted on the basis of the Katz-Postal (1964) analysis. Sentence constructions that are less closely related in the grammars (that is, sentences differing by more transformations or by more universal morphemes) are generally farther apart in the configuration. As was indicated by the Kruskal technique, the questions fall about midway between the affirmatives and the negatives. The distance between affirmative questions and negative questions is, as expected, very small, when compared to the affirmative-negative distance among nonquestions. In addition, the questions are somewhat removed from the plane of the

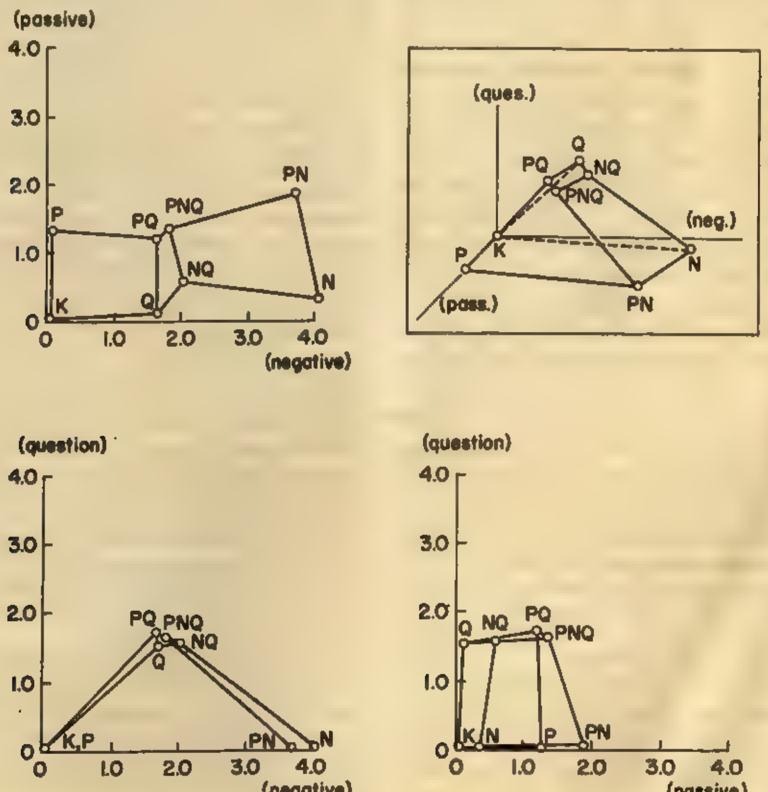

FIG. 4. Two dimensional projections of Experiment 1 configuration, with oblique projection inset.

nonquestions, a characteristic about which the Katz-Postal analysis seemed to make no prediction.

A passive-active distance comparable in magnitude to the question-nonquestion distance, and decidedly smaller than the affirmative-negative distance among nonquestions, appears in the configuration. As was expected, the passive-active distance among questions is comparable in magnitude to the passive-active distance among nonquestions, although perhaps slightly smaller in the former.

The configuration obtained may be said to be essentially that predicted on the basis of the Katz and Postal analysis, with the corrective that the distance corresponding to one constructional difference need not equal the distance corresponding to another. It was decided to see if the same general configuration would be obtained using a different technique for the measurement of the similarity of sentences.

Experiment 2

The second study was designed to obtain a measure of the confusions among the eight sentence constructions of a P,N,Q sentence family in a recognition task. The task used was derived from one used by Mink (1962) to study the generalization among associatively related words and is an extension of the procedure used by Clifton, Kurcz, and Jenkins (1965) to study confusion among the K,P,N, and PN sentence constructions. Essentially, *S* in this task is shown a list of sentences a small number of times, and then is shown a longer list, including sentences transformationally related to those of the first list, with instructions to press a telegraph key to every sentence that appeared on the first list.

Method

The method will be sketched in rough outline first. Each of 48 *Ss* was given six presentations of a list of eight sentences. Each sentence had a different content (was a member of a different P,N,Q sentence family) and was in a different one of the K,P,N,..., PNQ grammatical constructions. The *S* was then presented once with a list of 128 sentences, and instructed to press a telegraph key whenever he recognized a sentence from the first (training) list. Sixty-four of these 128 sentences were transformationally related to the training list sentences, and 64 were unrelated. Each sentence on the training list had eight sentences related to it on the second (test) list. These eight sentences were all the members of the sentence family of which the training list sentence was a member.

A measure of the confusions between any training list sentence and all the sentences related to it by some combination of the passive, the negative, and the question transformations could thus be obtained. These confusion measures could be presented in an 8×8 confusion matrix in which the rows represent the test sentence constructions and the columns represent the training sentence constructions. A number of training list forms were used for the purposes of balancing and replicating the design.

Materials. Sixteen kernel sentences were selected from those listed in Table 5 and divided into two sets, A and B, as indicated. The eight members of the P,N,Q sentence family corresponding to each of these kernels were determined, yielding a total of 128 sentences.

Sixteen eight-sentence training list forms were

TABLE 5

KERNELS OF FAMILIES IN EXPERIMENTS 2 AND 3

Set A	Set B
Betty read the book.	The dentist filled the tooth.[a]
The plumber dropped the pipe.[a]	The carpenter built the house.[a]
The doctor cured the patient.[a]	The clerk sold the shirt.[a]
The woman spanked the child.[a]	Tom washed the window.
Bob rang the bell.	John hit the ball.
The man closed the box.[a]	The boy broke the glasses.[a]
The professor gave the lecture.[a]	The dog chased the cat.[a]
The girl wrote the letter.[a]	His mother prepared the dinner.
The truck moved the piano.[a]	The horse ate the hay.[a]
The tractor pulled the plow.[a]	Mary saw the fish.
The student wrote the paper.	The barber cut the hair.[a]
Joe painted the wall.	The bus carried the people.[a]

[a] Sentence used in Experiment 2.

constructed. The eight A forms each contained one sentence from each of the sentence families of the Set A sentences, and the eight B forms one sentence from each of the Set B families. Each of the 16 forms contained 1 sentence in each of the K,P,N, . . . , PNQ constructions, and each of the 128 sentences appeared on just 1 training list form. The particular construction in which a representative of a specific sentence family occurred in a training list form was determined by an orderly progression, so that one of the Set A families, say, would be represented by K in the first A training list form, by P in the second, by N in the third, etc. Three randomizations were made of each training list form, the same randomizations being applied to each form.

A single test list which contained *all* 128 sentences used in the experiment was constructed. Sixty-four of these sentences were members of Set A sentence families, and 64 were members of Set B families. The test list could be paired with each one of the 16 training list forms. When the test list was paired with a Form A training list, the 64 Set A test list sentences were related to the training list sentences, while the 64 B test list sentences were unrelated and served as control sentences. The status of these two sets of 64 sentences as related or control sentences was reversed when the test list was paired with a *B* training list form.

The relationships between the sentences in any one training list and the related sentences in the test list can be represented in an 8 × 8 matrix in which the rows represent the test sentence construction and the columns the training sentence construction, similar to the matrix used in Experiment 1. Each cell in this matrix was filled by one test list sentence in any particular training form–test list pairing. Further, the various training list forms were constructed so that the sentence families were balanced over the cells. That is, every sentence family was represented in every cell under some pairing. Finally, the balanced 8 × 8 matrix was replicated over the two sets of sentences, A and B.

Three randomizations of the test list were made. One randomization was initially made, and the other two randomizations were derived from the first by a systematic rearrangement of the first, second, and third thirds of the initial randomized list of sentences, with rerandomization within the thirds. Each of these three randomizations was paired with each of the 16 training list forms, yielding a total of 48 pairings.

Apparatus. The lists of sentences were typed on memory drum tapes in capital letters, with all punctuation marks except the question mark eliminated. A Lafayette memory drum was used to present the sentences to *S*s. A telegraph key was mounted in front of the memory drum and connected to a 12-volt transformer-operated buzzer audible to both *S* and the experimenter (*E*). A tape recorder was used to present the instructions.

Subjects and procedures. Forty-eight undergraduates at the University of Minnesota were used as *S*s. The *S*s were obtained from the introductory psychology class and were given experimental credit counting toward their course grade for participating in the experiment. Each *S* heard instructions indicating that he was going to see a list of sentences and instructing him to press the telegraph key immediately after he silently read each sentence. The *S* was instructed to try to remember the sentences on the list so that he could recognize them later. The three randomizations of one training list were then presented twice (for a total of six times through the list of sentences) at the rate of one sentence every 4 seconds, with a 4-second interval between randomizations.

Immediately after the presentation of the final randomization, *S* heard instructions indicating that he was going to see a longer list containing the sentences he had seen on the first list as well as some others. He was instructed to press the telegraph key whenever he thought that he recognized a sentence, but not to press it when he was sure that he did not recognize a sentence. He was then given one of the three randomizations of the test list at a 4-second rate.

Three *S*s were randomly assigned to each training list form, and one *S* in each form was randomly assigned to each test list randomization. Half the *S*s were randomly assigned to one of the two *E*s (PO), and the remaining *S*s to the other *E* (CC).

Results

The basic data consist of the proportion of *S*s pressing to the test list sentences in each relationship classification. These data are presented in Table 6, in the matrix of test list sentence–training list sentence relationships. These recognition proportions index the amount of confusion which occurred between the various pairs of sentence constructions, except for the entries on the main diagonal, which indicate the proportion of correct recognitions.

To the right of the matrix is a column containing the proportion of presses made to the control sentences. As was reported by Clifton, Kurcz, and Jenkins (1965), the number of false recognitions of control sentences, relative to false recognitions of related sentences, is negligible. It is obvious that there is much more generalization, or confusion, among the members of a transformationally defined sentence family than between unrelated sentences.

Since the study was, in effect, replicated over the two sets (A and B) of sentences, it is possible to gauge the reliability of the

TABLE 6
PROPORTION OF POSSIBLE PRESSES TO TEST LIST SENTENCES, EXPERIMENT 2

Test list construction	Training list construction								
	K	P	N	Q	PN	PQ	NQ	PNQ	Control
K	.77	.50	.33	.58	.19	.35	.46	.23	.02
P	.60	.85	.40	.54	.52	.71	.44	.52	.03
N	.33	.21	.65	.27	.50	.33	.40	.38	.02
Q	.27	.29	.35	.60	.21	.38	.44	.29	.02
PN	.17	.27	.50	.27	.73	.46	.31	.46	.02
PQ	.25	.38	.27	.35	.44	.58	.48	.69	.02
NQ	.33	.29	.54	.60	.48	.52	.71	.56	.03
PNQ	.29	.46	.38	.40	.54	.56	.50	.79	.05

confusion measures. One balanced confusion matrix was obtained for each of the sets of sentences, and the 64 scores in one matrix were correlated with the 64 scores in the other matrix. The obtained correlation equaled +.72. Since this correlation is analogous to a split-half reliability coefficient, the Spearman-Brown correction was applied, yielding a corrected reliability coefficient of +.84. A correlation of this magnitude is reassuring evidence of the stability of the phenomenon and of its invariance over sentences of different content.

One might be concerned about the dependence of the results upon the particular test list randomizations used, especially since only three different randomizations were used. To determine if the randomization was a critical factor, a Test List Randomization × Test List Sentence Classification (the classifications in the matrix of Table 6, omitting the diagonal cells) × *Ss* ($3 \times 56 \times 48$) analysis of variance was run. The entries in the analysis were the numbers 0 (no press) or 1 (press). A significant effect of test list sentence classification was obtained (F $(55,2475) = 4.14$, $p < .001$), while the Fs corresponding to test list randomization and to the randomization by sentence classification interaction were less than 1. It appears that test list randomization is not a critical factor in the results obtained.

As in Experiment 1, the Kruskal scaling technique was used to construct a multidimensional scale of the distances among the eight sentence constructions. The basic confusion proportion data may be considered direct measures of the similarity of the sentence constructions. The data were made suitable for scaling purposes by dividing each proportion, p_{ij}, by the proportion on the main diagonal in the same row, p_{ii}. This conversion serves two purposes: It provides a correction for any tendencies to respond more to some sentence constructions than to others, tendencies which seem to be present in situations like the present one (Clifton, 1964; Odom, 1964). Second, it serves to make the entries on the main diagonal all equal to 1, indicating perfect similarity between a sentence and itself. Symmetric entries in the matrix of converted scores (p_{ij}/p_{ii} and p_{ji}/p_{jj}) were averaged for the sake of the stability of the scores, and a halfmatrix of averaged converted scores was obtained. This halfmatrix, without the main diagonal, was to serve as the input data for the scaling program. It is of interest to note that the values in this halfmatrix of similarities correlated −.73 with the values in the halfmatrix of comparative distances obtained from the sentence rankings of Experiment 1. This indicates a fairly close relationship between the judged similarity of sentences and the recognition confusions among them.

Scaling was attempted in both two and three dimensions, using both Euclidean and city block spatial models. A satisfactory spatial representation could not be obtained in two dimensions, the minimum stress proving to equal 13.3% when scaling took place in Euclidean space, and 7.4% in a city block space. In three dimensions, however, a configuration with satis-

TABLE 7

ROTATED CONFIGURATION, EXPERIMENT 2

Construction	Dimension		
	Passive	Negative	Question
K	0.00	0.00	0.00
P	1.10	0.00	0.00
N	0.42	1.75	0.00
Q	0.31	0.74	1.11
PN	1.54	1.59	−0.02
PQ	1.46	0.74	0.77
NQ	0.39	0.94	1.16
PNQ	1.44	1.14	1.29

factory stress was obtained for both the Euclidean and the city block spatial models. The stress for the Euclidean configuration equaled 1.4%, and the stress for the city block, 1.9%.

Upon examination, it became clear that the three-dimensional configurations did not have the regular character needed for the city block metric to be applicable to the lattice-metric Miller-Chomsky transformational cube of Figure 1. For this reason, and due to the ubiquity of the Euclidean model in other applications of multidimensional scaling, it was decided to present only the results of the Euclidean scaling.

The obtained Euclidean configuration was translated and rotated to an a priori determined position, as was done in Experiment 1. The coordinates of the K point were set at (0,0,0), the Y and Z coordinates of P were set equal to 0 with the X coordinate left free to vary, and the Z coordinate of N was set at 0 with the X and Y projections being unfixed. The resulting configuration of sentence constructions is presented in Table 7. The first dimension can be considered an active-passive dimension, the second an affirmative-negative dimension, and the third a nonquestion-question dimension.

The two-dimensional projections of the

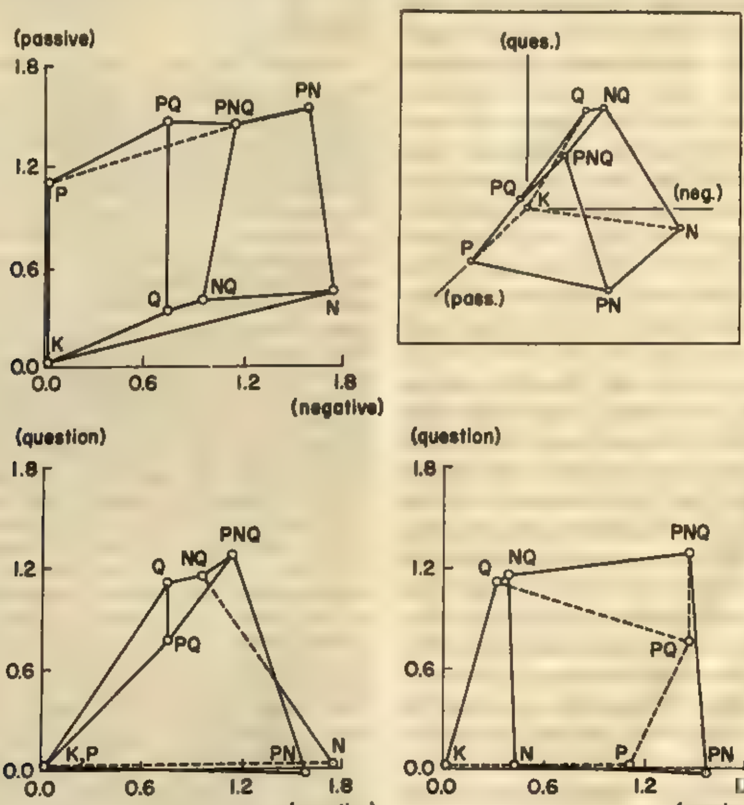

FIG. 5. Two dimensional projections of Experiment 2 configuration, with oblique projection inset.

configuration, as well as an oblique projection with the passive dimension foreshortened by $\sqrt{2}$, are presented in Figure 5. The configuration, it will be noted, bears a close resemblance to the configuration obtained in Experiment 1, using a metric scaling technique, except for the greater distance between passives and actives and some irregularity in the distances among the questions.

A number of conclusions may be drawn from an examination of the projections. As in Experiment 1, the grammatically more distantly related sentences are farther apart in the configuration. The questions fall about midway between the affirmative and the negative nonquestions. The Q point is very close to the NQ point, and the PQ point to the PNQ point. On the other hand, the distance between passive and active questions is sizable and comparable to the distance between passive and active nonquestions. Considering only the nonquestions, it appears that the distance between affirmatives and negatives is slightly greater than the distance between actives and passives. Clifton, Kurcz, and Jenkins (1965) found affirmatives and negatives to be significantly less similar (more distant from one another) than passives and actives, and a similar contrast, in exaggerated form, was obtained in Experiment 1. As in Experiment 1, the configuration obtained is essentially that expected on the basis of the Katz and Postal (1964) analysis.

For the purposes of the Kruskal scaling technique, the similarity measures (the converted response proportions) were considered to be measures on only an ordinal scale. However, for the purposes of the experiments to be reported next, it was thought worthwhile to consider similarity as being measured on an interval scale and to determine the function relating the similarity measures to the scaled distances. The relationship between converted response proportion, S, and distance, D, appeared to be continuous and linear (a test for significance of curvilinearity yielded an $F < 1$, with $df = 6/20$). The best-fit line (least squares) was described by the equation $D = 2.9 - 2.5\,S$.

Experiment 3

In the third study, confusions among subsets of the sentence constructions investigated in Experiment 2 were measured. The technique used was essentially the same as that in the previous study, except that each *S* was presented only four of the eight sentence constructions. Further, multiple examples of each of the four constructions—and thus of each of the 16 possible training sentence construction–test sentence construction relationships—were shown each *S*. Five groups of *S*s were run, each group receiving a different set of four syntactic constructions. The studies were designed to test the generality of the previously obtained measures of similarity among sentence constructions in a somewhat different experimental situation. Also, the technique of analysis used allowed a check on the generality of the function obtained in the previous study relating the similarity measures to scaled interpoint distance.

Method

Materials. The 24 K sentences listed in Table 5 were constructed, and divided into two sets, A and B. It will be noted that the sentences used include those used in Experiment 2. The seven other members of the P,N,Q sentence families of which the Table 5 kernels are members were determined.

A different set of training lists and test lists was constructed for each of five groups of *S*s. The lists differed among the groups only in the grammatical constructions of the sentences they contained. The Group 1 lists contained Q,PQ,NQ, and PNQ sentences; Group 2, K,P,Q, and PQ; Group 3, N,PN,NQ, and PNQ; Group 4, K,Q,NQ, and N; and Group 5, P,PQ,PNQ, and PN. It may be noted that these five groups of sentences form five of the six faces of the cube of Figure 1. The sixth face, K,P,N, and PN, was investigated by Clifton, Kurcz, and Jenkins (1965) and was also investigated in Experiment 4 of the present series.

The procedures followed in constructing the lists were those used by Clifton, Kurcz, and Jenkins (1965) and are very similar to those used in Experiment 2. A training list contained 12 sentences, 3 sentences in each of the four constructions. Eight training list forms were constructed for each group. Four forms contained one sentence from each of the families of the Set A sentences, and four contained one sentence from each of the Set B families. The four forms containing sentences from Set A families differed in the construction in which any particular sentence appeared, as did the four forms containing members of Set B families. The construction in which a sentence

appeared in a training list form was determined by an orderly progression, so that a sentence family that was represented by, for instance, a K in one training list form would be represented by a Q in the next form, an NQ in the next, and an N in the final form. Four training list randomizations were made. The same randomizations were applied to each form.

A single test list was constructed for each group. The test list for one group consisted of all the 96 sentence family members in the grammatical constructions presented to that group. Forty-eight of these sentences were members of Set A families, and 48 were members of Set B families. The test list was paired with each training list form received by the group for which it was designed. Thus, for each training list–test list pairing, 48 of the test list sentences were transformationally related to training list sentences, and 48 were unrelated. The training list sentence–test list sentence relationships for each group can be represented in a 4 × 4 matrix, with rows representing test list sentence construction and columns representing training list sentence construction. In Experiment 3, unlike Experiment 2, there were three examples of each of the 16 relationships: that is, 3 of the 48 related test list sentences could be classified in each cell of the matrix.

Three randomizations of the test lists were made, using the procedure described in Experiment 2. The apparatus used in that study was employed.

Subjects and procedures. Forty-eight University of Minnesota introductory psychology students were used in each group, for a total N of 240.

The procedures employed were basically the same as those of Experiment 2, and *S*s received essentially the same instructions. Each *S* was shown the four randomizations of a training list form, given the test instructions, and then shown one randomization of the test list appropriate for his group. Again, a recognition of a test list sentence was scored if *S* pressed a telegraph key when that sentence appeared.

The five groups were run successively, the first 48 *S*s being assigned to Group 1, the second 48 to Group 2, etc. Within each group, six *S*s were randomly assigned to each training list form. Each of the three test list randomizations was assigned to two of the *S*s receiving each training list form. One of the two experimenters (PO and CC) ran one of the *S*s receiving a particular training list form–test list randomization combination, and the other *E* ran the other *S* in that combination.

Method of Analysis

The basic data obtained from each group were the proportions of times *S*s responded to a test list sentence in a given training construction-test construction classification. Within each group, two independent sets of measures of these proportions were available, one from *S*s who received a training list containing the Set A sentences, and one from *S*s who received a training list containing the Set B sentences. These two sets of proportions could be correlated with one another to obtain an estimate of their reliability.

For the purposes of scaling, the obtained data were converted to distances using the formula that was found to relate converted similarity scores to the interpoint distances in Experiment 2. These distances were scaled using the metric technique employed in Experiment 1. The resulting configurations were compared with the corresponding subconfigurations of Experiment 2. For the purpose of facilitating this comparison, the interpoint distances of the Experiment 2 subconfigurations were scaled in the same fashion. Such a procedure results in configurations identical to the subconfigurations of the entire Experiment 2 configuration, Figure 5. The procedure was applied simply because it oriented these subconfigurations so that they would be directly comparable to the Experiment 3 configurations, and because it allowed a quantitative determination of the dimensionality of the subconfigurations.

Results

Reliability. The number of responses made to the Set A test sentences in each training sentence construction—test sentence construction classification was correlated with the number of responses made to the corresponding Set B test sentences for each group. The correlations, when corrected by the Spearman-Brown formula, ranged from +.88 to +.92 for the five groups. The correlations indicate generally satisfactory reliability of the confusion measures over two sets of sentences and subjects.

Scaling. The proportion of possible presses made to the test list sentences of the different classifications is presented in Table 8 for each of the five groups. The proportion of presses made to the control sentences is not shown, being uniformly quite low.

Converted similarity scores were calculated from the proportions of possible

TABLE 8

PROPORTION OF POSSIBLE PRESSES TO TEST LIST SENTENCES, EXPERIMENT 3

Group	Test list construction	Training list construction			
		Q	PQ	NQ	PNQ
1	Q	.62	.53	.59	.49
	PQ	.56	.67	.55	.72
	NQ	.72	.54	.72	.58
	PNQ	.58	.67	.66	.77
		K	P	Q	PQ
2	K	.77	.49	.45	.41
	P	.59	.85	.57	.83
	Q	.40	.35	.58	.44
	PQ	.33	.47	.43	.62
		N	PN	NQ	PNQ
3	N	.72	.44	.56	.35
	PN	.40	.69	.47	.59
	NQ	.61	.49	.76	.51
	PNQ	.47	.65	.59	.73
		K	Q	NQ	N
4	K	.70	.44	.33	.32
	Q	.38	.64	.51	.42
	NQ	.35	.58	.65	.51
	N	.25	.36	.38	.53
		P	PQ	PNQ	PN
5	P	.76	.60	.63	.51
	PQ	.51	.65	.53	.48
	PNQ	.53	.62	.73	.56
	PN	.31	.31	.41	.50

presses for each group by dividing each entry in the matrix by the main diagonal entry in its row, and the symmetrical converted scores were averaged. These averaged scores were converted to distances by the formula $D \doteq 2.9 - 2.5\ S$ (averaged converted similarity score), and the scalar products of the centroid-origin vectors of the configuration having these interpoint distances were determined. The resulting scalar product matrices were factored using the principal-axes procedure. The scalar product matrices of the interpoint distances of the subconfigurations obtained in Experiment 2 were also calculated and factored.

No sizable negative latent roots were obtained in any of the five analyses of the data of Experiment 3 or the five analyses of the Experiment 2 subconfigurations. The largest negative root was obtained in the analysis of the data of Group 1, Experiment 3, and it amounted in absolute value to 3% of the sum of the positive latent roots extracted. The Experiment 3 configurations thus may be considered to exist in real space.

The dimensionality of the configurations resulting from the factor analyses was tested in the same way as in Experiment 1. All the configurations could be considered as being two-dimensional without imposing too much distortion on the data. The sums of squares of the elements in the scalar product matrices which were factored were compared with the sums of squares of the elements of the scalar product matrices derived from the first two factors extracted in each analysis. The latter equaled 99.6% of the former for Group 1 of Experiment 3, and 99.5% for the corresponding subconfiguration of Experiment 2; 99.6% for Group 2 of Experiment 3 and 99.9% for the corresponding Experiment 2 subcon-

figuration; 98.9% for Group 3 of Experiment 3, and 99.7% for the corresponding Experiment 2 subconfiguration; 98.4% for Group 4 of Experiment 3, and 99.9% for the corresponding Experiment 3 subconfiguration; and 95.1% for Group 5 of Experiment 3 and 99.9% for the corresponding Experiment 2 subconfiguration.

The two-dimensional configurations of sentence structures were rotated and translated, and are presented in Figure 6. The configuration obtained for each group of Experiment 3 is placed next to the subconfiguration containing the same sentence structures in the Experiment 2 configuration.

While there are discrepancies within the pairs of corresponding configurations, it seems reasonable to emphasize the similarities between the Experiment 3 configurations and the corresponding Experiment 2 subconfigurations. First, all the Experiment 2 subconfigurations were essentially two-dimensional, with each dimension in each subconfiguration generally identifiable with a grammatical transformation. The configurations obtained by converting the Experiment 3 data to distances using the formula derived from the Experiment 2 results were also all essentially two-dimensional in the same fashion. Second, the most striking result of Experiment 2—the fact that affirmative questions and negative questions are relatively close together, about midway between the affirmative and the negative nonquestions—was obtained in Experiment 3. Third, the distance between affirmative and negative nonquestions is, in Experiment 3 as in Experiment 2, the greatest difference between any pair of sentences which may be analyzed as differing by a single transformation or a single universal morpheme. These three types of similarity between the results of the two experiments seem to confirm the dual assumptions (*a*) that similarities among sentence structures are essentially the same whether they are measured simultaneously among all eight structures investigated or are measured independently among subsets of these eight structures,

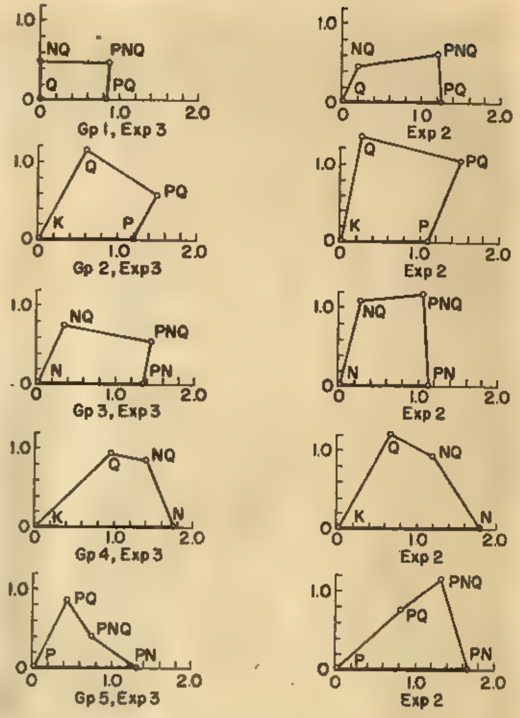

FIG. 6. Experiment 3 configurations, with Experiment 2 subconfigurations.

and (*b*) that the formula found in Experiment 2 to relate the measure of confusions among sentences to distances among them may be generalized to confusion measures obtained using a somewhat different procedure.

The discrepancies between the members of pairs of corresponding configurations should not be overlooked, however. The pattern of confusions among the four question constructions (Group 1, Experiment 3) was very similar to the pattern of confusions among the same four constructions obtained in Experiment 2. However, the distance between active questions and passive questions seemed to be smaller in the former than in the latter. That is, there were relatively more confusions between active and passive questions when the *S*s were shown only questions than when they were shown questions in the context of nonquestions.

The distance between P and PQ in Group 2 and the distances between N and

NQ and between PN and PNQ in Group 3 were relatively smaller than the corresponding distances obtained in Experiment 2. While an approximately equal number of confusions was found in Experiment 2 between sentences which could be analyzed as differing by the question transformation and between sentences differing by the passive transformation, generally more confusions were found between pairs of sentences of the former type than between pairs of the latter type in Experiment 3.

The configuration of sentences obtained in Group 4 showed no systematic deviations from the corresponding configuration of Experiment 2. However, the distance between P and PNQ in the configuration of Group 5 was noticeably small when compared to the P-PNQ distance in Experiment 2. It should be noted that the N-NQ distance in Group 4 and the P-PQ distance in Group 5 were not strikingly smaller than the corresponding distances in the Experiment 2 subconfigurations, while these same distances were relatively small in the Groups 2 and 3 configurations. However, the PN-PNQ distance in Group 5 did seem to be too small relative to the PN-PNQ distance of Experiment 2, just as it did in Group 3.

The extent to which these deviations may be considered reliable is not clear. There seems to be no particular system to the deviations, except for an inconsistent tendency for the various question constructions to be more often confused with each other and with nonquestion constructions when there is a smaller variety of different constructions with which they may be contrasted, as in Experiment 3. It does not seem, however, that this uncertain generalization may be extended to confusion with the K construction, which appears to be uniformly quite distinguishable from all other constructions in Experiment 3.

The results of Experiment 3, then, seem to confirm the major conclusions of Experiment 2. However, these results do differ from the Experiment 2 results in their fine details, for reasons that are not clear.

Experiment 4[4]

In Experiment 4 (as in Clifton, Kurcz, & Jenkins, 1965), the confusion among the K,P,N, and PN sentence constructions was investigated. The same basic procedures used in Experiment 3 were used. However, no control sentences appeared in Experiment 4, and the Experiment 3 technique was extended by giving repeated training and test trials on the sentences. Further, confusion among sentences containing the auxiliary verb "have" was compared with confusion among sentences not containing this auxiliary. The syntactical relationships among the K,P,N, and PN sentences remain the same whether or not the auxiliary appears in the sentence. However, the relative degree of graphemic or phonemic (or physical) similarity among the constructions changes appreciably when the auxiliary is added. The study should therefore shed a little light on the importance of physical resemblance for similarities among sentences.

Method

Materials. One set of 12 K sentences was selected from the sentences used in Experiment 3. The P,N, and PN forms corresponding to these kernels were determined. Four training list forms were constructed in the same fashion as the training lists of one sentence set were constructed in Experiment 3. Nine randomizations of each of these training list forms were made.

Six randomizations of a single test list were made. This test list contained all 48 sentences used, but did not contain any unrelated (control) sentences.

A parallel set of training and test lists was made, using sentences which contained a form of the auxiliary "have" (e.g., "John has hit the ball," "The ball has been hit by John," "John hasn't hit the ball," "The ball hasn't been hit by John"). These lists were identical to the first set of lists, except that all the sentences were in the present perfect construction rather than in the simple past.

Apparatus. The apparatus used was similar to that used previously, except that two memory drums were used, one for the training lists and one for the test lists.

[4] The data reported as Experiment 4 were collected by David M. Harrington and Michael Ryan while they were NSF Undergraduate Summer Fellows at the University of Minnesota. The authors express their gratitude to them for allowing the use of their data in this report.

Subjects and procedures. Two groups of 24 University of Minnesota summer session students were run. Twelve *S*s in each group were male, and 12 were female. The Group 1 *S*s were assigned to the lists containing sentences in the simple past, and the Group 2 *S*s to the lists containing present perfect sentences. Six *S*s in each group were assigned to each training list form, and two *S*s receiving one training list form were assigned to each of three orders of presentation of the test list randomizations. Each *E* (DH and MR) ran three *S*s assigned to each of the training list forms in each group, and each *E* ran one-half male and one-half female *S*s. With these restrictions, *S*s were randomly assigned to conditions and *E*s. The *S*s were given the instructions used in Experiment 3, modified to explain the use of two memory drums and to describe the repeated trials procedure used. An abbreviated set of instructions were read the *S*s after each exposure of the test list.

Each *S* was shown four randomizations of the training list form to which he was assigned, at a 4-second rate. He was then moved to the other memory drum and was shown the initial randomization of the test list. Upon completing the first test list randomization, *S* was moved back to the first memory drum, briefly reinstructed, and shown a new single randomization of his training list form. He then returned to the second memory drum and was shown a second test list randomization. The cycle was repeated until *S* had received six different test list randomizations. A 3-minute rest was given between the third test list and the fourth training list.

Results

The proportion of possible presses to test list sentences on the first and sixth trials is shown in Table 9 for Group 1, and in Table 10 for Group 2. It is of interest to note that the values for the first trial of Group 1 correlate +.95 with the corresponding values reported by Clifton, Kurcz, and Jenkins (1965), using a different sample of *S*s, a somewhat different set of sentences, and 48 unrelated control sentences on the test list. It may be noted that there appeared to be better discrimination (more responding to training sentences, less to related test sentences) in the present study than in the study by Clifton et al. This may reasonably be attributed to the lack of any distracting control sentences in the present study. Still, the pattern of generalization is remarkably constant across the studies.

Also worthy of note is the fact that the first trial scores in Table 9 (Group 1, simple past sentences) correlated +.95 with the first trial scores in Table 10 (Group 2, present perfect sentences). There does seem to be a higher level of generalization among the sentences seen by Group 2 than those seen by Group 1. A Tense × Generalization Category (off-diagonal cell in the matrix) × Trials analysis of variance supports this impression. A significant effect of tense (F (1, 46) = 5.78, $p < .05$) was found, as well as a significant effect of category (F (11, 506) = 38.85, $p < .01$). A significant trials effect was also found (F (5, 230) = 58.72, $p < .01$). Further, the Tense × Categories and the Categories × Trials interactions were significant (F (5, 506) = 3.08, $p < .01$, and F (55, 2530) = 2.31, $p < .01$, respectively). The Tense × Categories × Trials interaction approached, but did not reach, significance (F (55, 2530) = 1.33, with 1.42 needed for the .05 level). The Tense × Trials interaction did not approach significance ($F < 1$).

The tense effect indicates that there was more generalization among present perfect sentences than among simple past sen-

TABLE 9

PROPORTION OF POSSIBLE PRESSES TO TEST LIST SENTENCES, EXPERIMENT 4, SIMPLE PAST SENTENCES (GROUP 1)

Test list construction	Training list construction							
	Trial 1				Trial 6			
	K	P	N	PN	K	P	N	PN
K	.90	.43	.21	.08	.92	.17	.00	.04
P	.42	.79	.28	.32	.10	.88	.01	.06
N	.17	.12	.72	.49	.08	.09	.93	.31
PN	.03	.19	.57	.75	.03	.14	.24	.96

TABLE 10

PROPORTION OF POSSIBLE PRESSES TO TEST LIST SENTENCES, EXPERIMENT 4, PRESENT PERFECT SENTENCES (GROUP 2)

Test list construction	Training list construction							
	Trial 1				Trial 6			
	K	P	N	PN	K	P	N	PN
K	.72	.49	.29	.15	.89	.21	.14	.04
P	.57	.75	.24	.32	.32	.90	.07	.12
N	.29	.26	.74	.56	.17	.10	.88	.28
PN	.25	.30	.61	.81	.07	.12	.35	.92

tences, presumably because of the greater physical similarity of the former. The categories effect indicates that certain pairs of constructions were more often confused than other pairs of constructions. The trials effect simply indicates that generalization decreased over trials, and the Trials × Categories interaction indicates that generalization decreased differentially for the various categories. The graph in Figure 7 may help to clarify this interaction. Here the various categories of sentences were combined into three categories, namely test list sentences whose construction differed from their training list construction by just the passive transformation (Pass), test sentences that differed by just the negative transformation (Neg), and test sentences that differed by both transformations (Pass + Neg). The proportion of possible presses made to the 12 sentences in each of these cate-

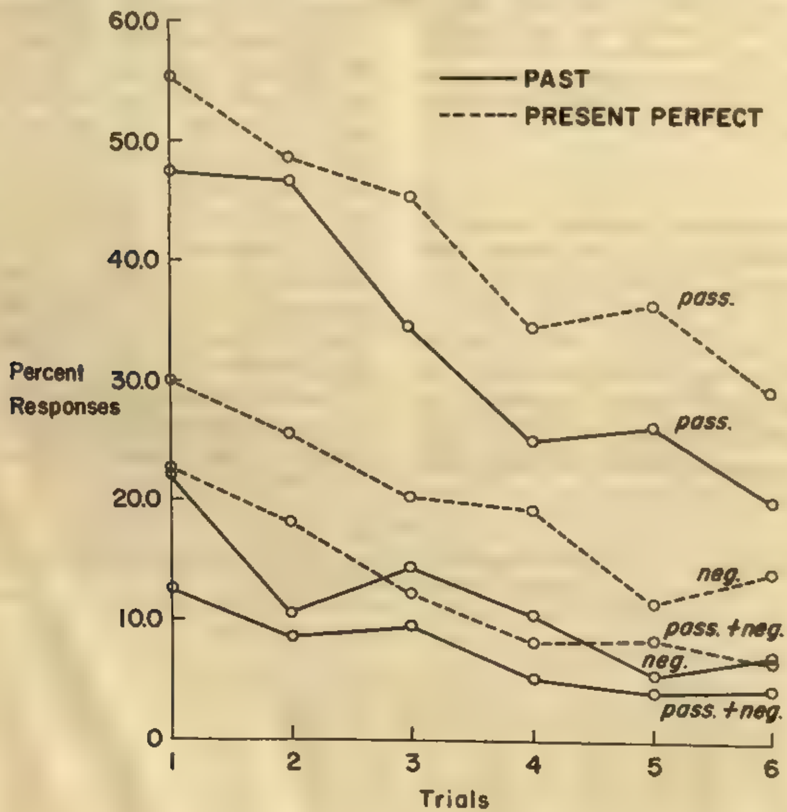

FIG. 7. Percentage of possible responses to test sentences differing from training sentences.

gories is plotted against trials. It seems that the relative number of erroneous recognitions of test sentences differing from training sentences by just the passive transformation decreased more rapidly with trials than did the relative number of erroneous recognitions of sentences in the other categories. This may simply reflect the fact that the test list sentences which differed from training list sentences by the negative transformation reached an asymptotically low level of responding in the later test trials.

The significant Tense × Categories interaction seems to indicate that the pattern of generalization differed between the simple past and the present perfect sentences. An examination of the scaled data, presented next, will aid in the interpretation of this interaction.

Scaled data. The data from each trial of each group were treated in precisely the same fashion as were the data from each group in Experiment 3. The interpoint distances among the K,P,N, and PN constructions obtained in Experiment 2 were also scaled in the same way as the Experiment 2 data were scaled in the Experiment 3 analysis. All the resulting configurations could be said to exist in real space. The largest negative root was obtained in the data of Group 2, Trial 5, and its absolute value was equal to only 0.2% of the value of the sum of the positive roots.

The K-P-N-PN subconfiguration from Experiment 2, and the configurations obtained in the early trials of Experiment 4, were essentially two-dimensional. However, the configurations of the data from the later trials of Experiment 4 do not seem to be two-dimensional. It becomes more and more necessary to use a third dimension in the representation of the configurations of the later trials. For the subconfiguration of Experiment 2, the sum of squares of the scalar product matrix derived from the first two factors extracted equaled 100.0% of the sum of squares of the original scalar product matrix which was factored. For Group 1 of Experiment 4, the corresponding values were: Trial 1, 99.9%; Trial 2, 98.8%; Trial 3, 96.7%; Trial 4, 89.1%; Trial 5, 90.4%; and Trial 6, 86.0%. For Group 2 of Experiment 4, the values were: Trial 1, 99.9%; Trial 2, 99.6%; Trial 3, 99.7%; Trial 4, 97.0%; Trial 5, 95.1%; and Trial 6, 91.2%. This increase in the dimensionality of the configurations over trials is reflected in the Trials × Categories interaction obtained in the analysis of variance.

The translated and rotated two-dimensional representations of the configurations obtained in Experiment 4, and of the Experiment 2 subconfiguration, are presented in Figure 8. The third dimension is represented by vectors attached to the labeled points in the two-dimensional configurations. The configurations were not rotated or translated in the third dimension. The length of each vector equals the value of the coordinate of each point in the third dimension; a vector extending

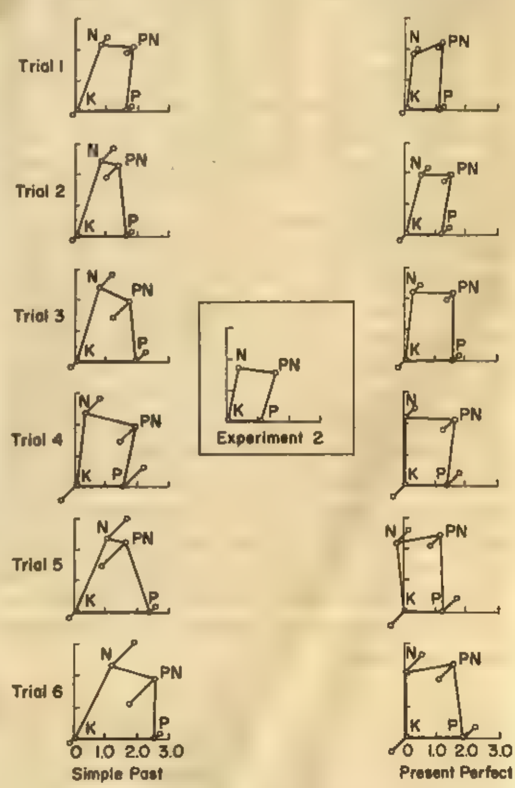

FIG. 8. Experiment 4 configurations, Trials 1-6, with Experiment 2 subconfiguration.

to the left and down from the point indicates a positive value of the projection, and a vector extending to the right and up indicates a negative value.

The sizes of the configurations may be compared across trials and across groups. Even ignoring the growth of the third dimension, there is an orderly growth in the size of the configurations over the trials, indicating the smaller number of confusions made in the later trials. Further, the configurations of the sentences shown Group 2 are generally smaller than the configurations of the sentences shown Group 1, indicating the greater number of confusions made among present perfect sentences than among simple past sentences.

There is a good deal of similarity in shape among the two-dimensional representations of the configurations. A passive dimension and a negative dimension are identifiable in each configuration. Also, the distance between negative and affirmative is greater in each configuration than is the distance between passive and active. These similarities confirm the conclusion reached in Experiment 3 that very much the same similarity relations among sentences are being measured in the different variations of the recognition technique. Further, the similarities between the configurations of Groups 1 and 2 in Experiment 4 indicate that the similarity relations being measured are, at least, not entirely dependent upon the formal (physical) similarities among the sentences, since Group 1 sentences were in the simple past while the Group 2 sentences were in the present perfect. This difference in tense markedly changes the physical relationships among the sentences (contrast "The boy hit the ball"; "The boy didn't hit the ball" with "The boy has hit the ball"; "The boy hasn't hit the ball") while leaving the underlying grammatical relationships unchanged.

As in Experiment 3, however, there are differences among the configurations which should not be overlooked. The greater size of Group 2 configurations has already been commented upon. Upon examination, at least one further difference becomes apparent. In the Group 1 configurations, the distance between K and P is almost always greater than the distance between N and PN. Clifton et al. (1965) also found a greater (though not significantly greater) distance between K and P than between N and PN. In the Group 2 configurations, this difference in distances is minimal and less consistent than in the Group 1 configurations. It is possible that the significant Tense × Sentence categories interaction reported earlier reflects the variation between groups in this difference. It is interesting to note that the Experiment 2 subconfiguration is more similar in shape to the Group 2 configurations than to the Group 1 configurations in Experiment 3, even though the sentences in the latter configurations were in the same tense as the Experiment 2 sentences.

A consideration of the possible reasons for these discrepancies and for the growth in the third dimension over trials, will be reserved for the discussion. However, before the discussion, we will present a reanalysis of data reported by Mehler (1963). Mehler's study further extends the range of techniques with which essentially the same similarity relationships among the sentence constructions being investigated are obtained.

The Mehler Experiment

Mehler (1963) reported the results of an experiment in which he studied the syntactic errors in the free recall of sentences. The reader is referred to Mehler's report for a complete description of the procedure. Essentially, Mehler presented his *Ss* with a group of the K,P,N, . . . , PNQ sentence constructions, each sentence representing a different sentence family. All his sentences were in the present perfect tense, for example, "The man has bought the house" and "Hasn't the secretary typed the paper?" The list of sentences was presented for five trials, the *S* being requested to recall all the sentences he had heard after each trial.

Mehler scored the sentences which were

recalled as correct (if identical to a presented sentence, or differing only in tense, in the replacement of the definite article by the indefinite article, or in the replacement of a word by a synonym), as syntactically erroneous (if the recalled sentence was a different member of the same P,N,Q sentence family as a presented sentence), or as otherwise erroneous.

Correctly recalled sentences and syntactically erroneous sentences can be classified in the 8 × 8 sentence construction by the sentence-construction matrix used in Experiments 1 and 2. The classification of a recalled sentence depends on the construction in which it was presented (the rows in the matrix) and the construction in which it was recalled (the columns in the matrix). The frequencies with which recalled sentences fall into each classification can be entered into the cells of this matrix. Mehler presented just such a matrix as Table 1 of his article.

The frequency measures in such a matrix may be considered to be similarity measures, and scaled using the Kruskal technique. However, it seemed wise to apply certain corrections to the data before scaling. First, since not every sentence presented was recalled correctly or with only a syntactic error, the row frequencies in the frequency matrix (the frequencies with which sentences presented in a given construction were recalled, regardless of the construction in which they were recalled) were unequal. Therefore, the frequency of f_{ij} in each cell was divided by the appropriate row frequency $\sum_{j=K}^{PNQ} f_{ij}$ to obtain proportion measures p_{ij} for each cell. Second, there seemed to be tendencies to recall sentences in certain constructions more than in others, regardless of the constructions in which the sentences were presented, for example, there was a definite tendency toward recalling sentences in the K form. That is, there were differences among the column frequencies. Therefore, each p_{ij} entry was divided by the main diagonal entry in the same *column*, p_{jj}, yielding corrected confusion values, c_{ij}. This particular correction was used in order to make the entries on the main diagonal equal to 1, indicating perfect similarity between any sentence and itself. Finally, the symmetrical corrected confusion values c_{ij} and c_{ji} were averaged, yielding a halfmatrix of averaged corrected confusion values (Table 11).

The entries in this halfmatrix were correlated with the corresponding halfmatrix entries of Experiments 1 and 2. The correlation between Mehler's mean c_{ij} values and the Experiment 1 comparative distances proved to equal −.66, and the Mehler values—Experiment 2 converted scores correlation equaled +.83. The magnitude of these correlations, in particular the latter correlation, is surprising when it is recalled that Mehler's sentences were of different content than the sentences in the other studies and that the former and not the latter contained a form of the auxiliary verb "have." It again appears that the various measures of sentence similarity are tapping the same, very stable, syntactic property of sentences.

The halfmatrix of averaged corrected confusion values, less the main diagonal, was scaled with the Kruskal technique. As was the case in Experiment 2, two-dimensional solutions were less than satis-

TABLE 11

AVERAGE CORRECTED CONFUSION VALUES: MEHLER

	K	P	N	Q	PN	PQ	NQ
P	0.097						
N	0.091	0.013					
Q	0.081	0.021	0.081				
PN	0.020	0.065	0.135	0.028			
PQ	0.027	0.155	0.031	0.108	0.095		
NQ	0.068	0.038	0.097	0.232	0.044	0.091	
PNQ	0.011	0.070	0.006	0.067	0.092	0.294	0.167

TABLE 12
ROTATED CONFIGURATION: MEHLER

Construction	Dimension		
	Passive	Negative	Question
K	0.00	0.00	0.00
P	1.33	0.00	0.00
N	−0.02	1.33	0.00
Q	0.00	0.69	1.14
PN	1.32	1.35	−0.08
PQ	1.33	0.70	1.14
NQ	0.00	0.69	1.14
PNQ	1.34	0.70	1.14

factory. With the Euclidean model, the best fitting configuration had a stress of 17.1% while the best city block configuration had a stress of 8.6%. In three dimensions, excellent fits were obtained in both Euclidean and city block spaces. The best fitting Euclidean configuration had a stress of 0.0%, while the best fitting city block configuration had a stress of 2.1%. For precisely the reasons given in Experiment 2, only the Euclidean configuration will be considered here.

The zero stress Euclidean configuration was translated and rotated to the position of the configuration obtained in Experiment 2. The coordinates of the obtained configuration are presented in Table 12. As in Experiment 1 and Experiment 2, it is possible to label the dimensions "passive," "negative," and "question," and once again it is noted that the questions fall about midway between the affirmative nonquestions and the negative nonquestions.

The two-dimensional and oblique projections of the configuration are shown in Figure 9. It is apparent that the obtained configuration is precisely the regular figure that our interpretation of the phrase-structure analysis implied. The points corresponding to affirmative questions and nega-

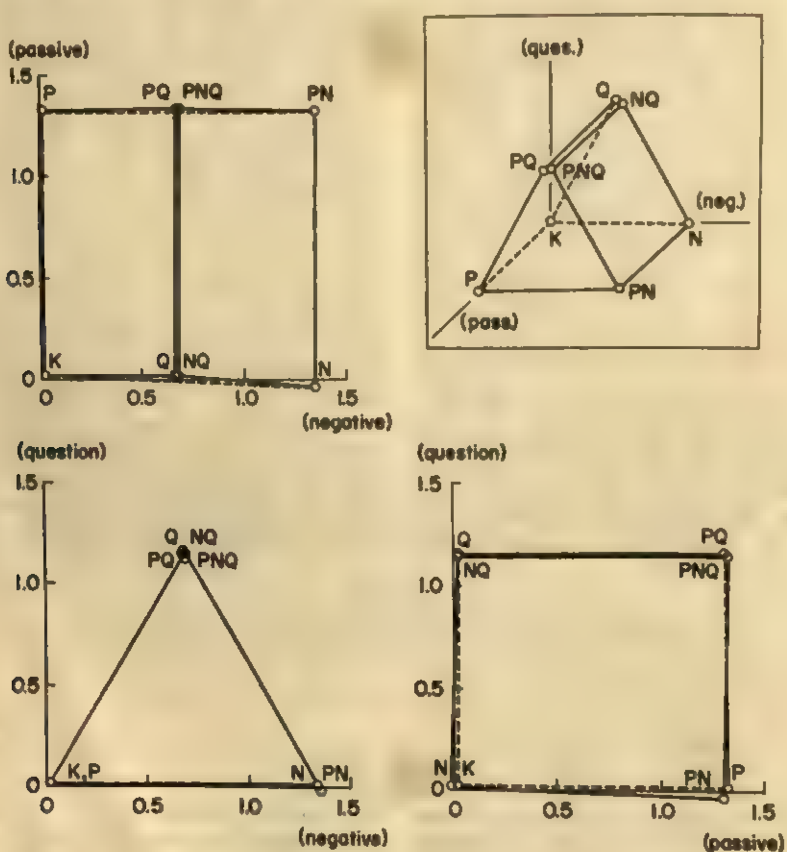

FIG. 9. Two dimensional projections of Mehler configuration, with oblique projection inset.

tive questions have almost the same location in the configuration, while the distance between active and passive questions is approximately the same as the distance between active and passive nonquestions. The questions are midway between affirmative and negative nonquestions, and the distances between passive and active, and between question and nonquestion, are approximately equal to the distance between affirmative and negative nonquestions.

One possible difficulty with the obtained scaling becomes apparent upon examination of a plot of distance values in the configuration against the similarity measures. As in Experiment 1, this plot is a "step function." Here, similarity values within one of three rather wide ranges of value are associated with one of three values of distance. Apparently, in the Mehler data, there were two clusters of sentence constructions (Q and NQ; PQ and PNQ) in which the constructions were very close to each other, relative to their proximities to the remaining constructions, while the remaining constructions were well separated from each other.

The "step function" obtained might lead one to feel that, in the present analysis, not all the information in the data is being used in the construction of the spatial configuration. However, whether or not it is thought that a more powerful treatment of the data would provide a more refined estimate of the true similarities among the constructions, it seems safe enough to accept the gross implications of the configuration obtained. The conclusion that the configuration supports our tentative interpretation of the Katz-Postal phrase-structure analysis of syntactic relationships is inescapable.

DISCUSSION

In the studies reported here, a variety of measurement techniques—techniques for determining judged similarity, recognition errors, and reproduction errors—were applied to the problem of ascertaining psychological similarity among sentence constructions. The sentence constructions investigated were those which were members of P,N,Q sentence families. In experiments using a recognition technique (Experiments 2 and 3) much more confusion was obtained among members of one sentence family than between members of different sentence families, indicating some relatively close relationships among sentences which are grammatically related. Of greater interest, a consistent *pattern* of similarity relationships was found among the members of a sentence family. Very similar patterns were found in the three experiments that simultaneously investigated all eight members of a sentence family (Experiments 1, 2, and the Mehler study). When separate subsets of the sentence constructions were examined in Experiments 3 and 4, relationships were obtained that were similar to those obtained in the other studies, although certain deviations were noted.

Multidimensional scaling techniques were used to analyze the patterns of relationships obtained. These scaling techniques resulted in multidimensional configurations of sentence constructions, where distance in a configuration was related to the dissimilarity of the sentence constructions. Each configuration could be dimensionalized in terms of the manner of grammatical relatedness of the sentence constructions. That is, one dimension along which sentence constructions were displaced from one another corresponded to the difference between active and passive sentences, another to the difference between affirmative and negative sentences, and a third dimension to the difference between nonquestion and question sentences. With certain exceptions in the case of the question sentences, the dimensions along which two constructions in the configuration were displaced from each other were those identified with the particular grammatical characteristics differentiating the sentences. Thus, grammatically less closely related constructions were generally further apart in the configuration than more closely related constructions.

Deviations from the dimensionalization in terms of apparent grammatical differences appeared among the questions. Affirmative questions and negative questions were consistently close together in the

configurations. Further, affirmative questions and negative questions all had approximately the same projection on the affirmative-negative dimension, falling about half-way between affirmative and negative nonquestions on this dimension.

The obtained relationships involving questions were not those expected on the basis of a transformational grammatical analysis of the constructions involved (see Figure 1 for a graphical presentation of these implications). However, they were very nearly those expected on the basis of an interpretation of the Katz and Postal (1964) phrase-structure analysis of the constructions (Figure 2).

Another characteristic in which the configuration obtained deviated from that which might be expected on the basis of the transformational analysis was the generally greater size of the deviations along the affirmative-negative dimension, relative to the size of the deviations along the active-passive dimension. The passive transformation is a more complex transformation than the negative transformation (in terms of the number of symbols in the string being transformed that undergo change, i.e., the number of elementary transformations involved). One might thus expect to find a greater perceived difference between sentences differing by the passive transformation than between sentences differing by the negative transformation. Such was not the case. The Katz and Postal analysis seemed to make no prediction regarding the relative sizes of the deviations along the dimensions.

Finally, it may be pointed out that the configurations appeared to be better described in a Euclidean space, as implied by the Katz and Postal analysis, than in the non-Euclidean space indicated by the transformational analysis. However, the scaling techniques used do not provide a really satisfactory basis for choosing the better of the two spatial models.

One might ask, does the pattern of similarity among sentences really reflect the grammatical relationships among the sentences? Might it not instead be reflecting the relationships among the sentences in meaning, or simply in phonetic or graphemic (physical) similarity? The question is not easily answered. One might point to the success achieved by the grammatical analysis. Such success is not presently possibly by an analysis in terms of meaning, simply because no method exists for specifying the relative closeness in meaning of different sentence constructions. However, certain points in the data, such as the greater deviations along the affirmative-negative dimension than along the active-passive dimension, seem likely to be congruent with a meaning analysis. Also, it should be remembered that the Katz and Postal (1964) phrase-structure analysis was designed with an eye toward an analysis of the meaning of sentences. One could work toward a determination of the role of meaning by investigating sentence constructions which are grammatically related in ways other than those investigated to see if predictions made on the basis of grammatical relationships continue to hold. In addition, it would be of interest to investigate sentences which appear to be related semantically but not grammatically, for example, sentences in which words are replaced by synonyms or by opposites.

Unlike the case of semantic similarity, there is no lack of ways in which to specify the physical similarity of sentences. In fact, the problem here lies in choosing among the many alternatives. Rather than attempting to defend the choice of any particular measure of physical similarity, however, we shall simply point out two aspects of the data which seem to indicate the insufficiency of an explanation in terms of physical similarity.

First, the obtained pattern of perceived similarities was much the same, whether simple past or present perfect sentences were investigated (contrast Experiments 1 and 2 with the Mehler study, and Group 1, Experiment 4 with Group 2, Experiment 4). The grammatical relationships among the sentence constructions remain the same regardless of the tense of the sentences, while the physical relationships change markedly. Among the simple past sen-

tences, for instance, the physical difference between affirmative and negative nonquestions seems to be far greater than the difference between affirmative and negative questions (compare "John hit the ball"; "John didn't hit the ball" with "Did John hit the ball?"; "Didn't John hit the ball?"). However, it is hard to find a comparable physical contrast among the corresponding sentences in the present perfect ("John has hit the ball"; "John hasn't hit the ball" versus "Has John hit the ball?"; "Hasn't John hit the ball?"). Nevertheless, it was consistently found that the questions are psychologically very similar, while the nonquestions are very dissimilar, regardless of the tense of the sentence.

Second, there are cases in which the constructions of one pair of sentence constructions are judged to be more similar, or are more frequently confused, than the constructions of a second pair, while the constructions of the second pair are quite obviously more similar physically than the constructions of the first pair. For instance, the active and the passive are generally perceived as being more similar than the affirmative and the negative, while it seems clear that the affirmative and the negative are *physically* more similar than the active and the passive (compare "The man closed the box"; "The box was closed by the man" with "The man closed the box"; "The man didn't close the box").

This is not to say that physical similarity is of no importance in the present results. The greater number of confusions among sentences in the present perfect (Group 2, Experiment 4) than among sentences in the simple past (Group 1, Experiment 4) is consonant with the apparently greater physical similarity among present perfect sentences. Also, the fact that N and PN appeared to be closer together than K and P among simple past tense sentences, but not among the present perfect sentences (Experiment 4), may reflect an effect of physical similarity. (It is difficult, however, to point to any aspect of physical similarity which would account for this difference, aside from sentence length.) Finally, it is possible that the deviations between the relationships obtained in Experiment 3 and the corresponding relationships obtained in Experiment 2 may reflect some kind of an interaction between physical similarity and the context in which the sentences are seen. The sentence constructions which were perceived as relatively more similar in Experiment 3 than in Experiment 2 are generally sentences which differ only in word order (e.g., P versus PQ, "The cat was chased by the dog" versus "Was the cat chased by the dog?" and PN versus PNQ, "The pipe wasn't dropped by the plumber" versus "Wasn't the pipe dropped by the plumber?"). Perhaps the identity of the individual words in these constructions assumes an important role in determining the number of confusions made when a smaller variety of sentence constructions is seen in the context of the experiment.

It will be recalled that, in Experiment 4, it was necessary to consider the configurations obtained in the later trials as being three-dimensional. That is, there was a growth over trials in the size of a third dimension which did not reflect any grammatical characteristic of the sentences. In effect, the sentence constructions became more nearly equidistant in the later trials. One might speculate that, in these later trials, *S*s are no longer discriminating on the basis of grammatical structure, but have selected certain unique characteristics of each sentence and are reacting to these characteristics. That is, whatever confusions occur among sentences on the later trials are traceable to similarities among certain (possibly physical) characteristics of the sentences, rather than to the grammatical relationships among the sentences. Alternatively, one might simply say that a minimum level of confusions is being approached in the later trials for all sentence constructions, implying that all constructions must appear to be equidistant from one another. Finally, one might suggest that what exists here is a scaling problem, where the function relating measured similarity to distance in the configuration is

really, say, exponential, and only appeared to be linear in the range of similarity values obtained in Experiment 2.

Further points could be discussed. For instance, it has been hypothesized (Mehler, 1963; Miller, 1962) that K construction is somehow central, that is, that it is basic to the other constructions. (A more proper statement would be that the terminal string underlying the K is basic to the strings underlying the other constructions.) Such an hypothesis is consonant with a transformational analysis of sentence relationships, but not with the analysis that indicates the different sentence constructions to be derived by different phase-structure rules. In the present analyses, it did not appear that the K construction was in any special way distinguished from the other constructions. Actually, any distinction that the K might have had might have been obscured by the corrections for response bias employed in the analyses. In Experiments 2, 3, and 4, this correction took the form of an adjustment for tendencies to "recognize" sentences in some constructions more than in others (although it should be noted that there seemed to be no bias toward "recognizing" K sentences: Clifton, 1964; Odom, 1964). In the analysis of the Mehler study, the correction adjusted for a tendency which did exist to recall sentences in the K form. It might be suggested that K is distinguished from the other forms not on any grammatical basis, but simply on the basis of its greater frequency of use in the language or perhaps its shorter length, and thus that the application of the corrections was legitimate for our purposes.

Finally, the topic of the relation between a generative grammar and the linguistic abilities of a speaker could be discussed. However, the present data justify no new strong assertions about this relationship. It is perhaps sufficient to point out that the studies reported here give substantial evidence for the existence of some parallel between the linguistic description of a language and the reactions of a language user to his language.

SUMMARY

Certain aspects of modern generative grammars were discussed. The implications of two types of grammars for syntactic relationships among sentence constructions were examined. One type of grammar treats certain related constructions as being transformational variants of a single "terminal string," and thus as being related to each other by sets of transformations. This type of grammar indicates that certain sentence constructions would be related to each other by an inverse function of the number of transformations by which they (or better, their underlying strings) differ. This type of relationship may be represented graphically in the cube of Figure 1, for sentences formed using some combination of the passive, the negative, and the question transformations.

Another type of grammar, exemplified by the grammar presented by Katz and Postal (1964), treats apparently related sentences as being related by virtue of their phrase-structure derivations. Specifically, it can be argued that sentence constructions are syntactically related if their underlying strings differ only in "universal" morphemes and that the closeness of the relationship is an inverse function of the number of universal morphemes by which these strings differ. It was argued that this phrase-structure treatment indicates grammatical relationships among sentence constructions which are similar to those indicated by the transformational approach, with certain important exceptions. Specifically, in the case of the sentence constructions presented in Figure 1, the Q and NQ, and the PQ and PNQ, are very similar to each other, and questions in general are approximately equally closely related to affirmative and negative nonquestions. These relationships were presented graphically in Figure 2.

A series of studies was carried out to determine empirically the perceived similarity of certain sentence constructions. Each study yielded a matrix of similarity

or dissimilarity measures, in which the rows and columns referred to sentence constructions. These matrices were analyzed using multidimensional scaling techniques, either metric techniques described by Torgerson (1958) or a nonmetric technique proposed by Kruskal (1964a, 1964b). These analyses resulted in multidimensional spatial configurations which could be compared with those predicted by the two grammatical analyses.

The first study used a judgment technique (method of multidimensional rank order) to determine the similarity of eight sentence constructions (K, P, N, Q, . . . , PNQ). The resulting data did not scale in a satisfactory manner when the Kruskal nonmetric technique was used. Apparently, subsets of the sentence constructions were more similar to one another than they were to any constructions outside the subsets, a property of the data that results in an excessively low dimensionality of the scaled configurations. However, when the data were scaled using the metric technique, a three-dimensional configuration emerged that was very similar to the one indicated by the Katz and Postal phrase-structure analysis, with the corrective that passives and actives were highly similar to each other.

The second study investigated the same eight constructions, using a recognition task, in which the confusions between related sentences in different constructions indexed the similarity of the constructions. When these data were scaled using the Kruskal technique, a very satisfactory configuration was obtained. This configuration again was much like that predicted by the phrase-structure analysis, with certain irregularities of dubious reliability among the question constructions.

In the third experiment, the sentence constructions investigated in Experiment 2 were examined four at a time, using the recognition technique. The resulting confusion data were converted to distances using the formula that was found to relate the confusion measure to distance in Experiment 2, and these distances were scaled using the Torgerson (1958) technique. The five two-dimensional configurations which emerged were compared with the corresponding subconfigurations of the Experiment 2 configuration. Important similarities were found, namely, each configuration could be dimensionalized in terms of the phrase-structure relationships among the constructions; affirmative and negative questions were again very similar to each other and midway between affirmative and negative nonquestions; and the distance between affirmative and negative nonquestions was, as before, the greatest distance between constructions differing by only one transformation or one universal morpheme. However, there were certain differences between the Experiment 3 configurations and the corresponding Experiment 2 subconfigurations. While these differences were of uncertain reliability and did not fit perfectly into any summarizing pattern, the best description of them would seem to be that sentences that differ only in word order are more often confused with each other when presented in the context of a smaller (Experiment 3) rather than a larger (Experiment 2) variety of sentence constructions.

In the fourth experiment, confusions among another subset of four sentence constructions (K, P, N, and PN) and the changes in confusions over repeated training and test trials were investigated. Further, confusions among sentences in the simple past tense were compared with confusions among sentences in the present perfect. The data were analyzed as in Experiment 3. On the first trial, the obtained configurations were much the same as the corresponding subconfiguration of Experiment 2. There were generally more confusions between sentences in the present perfect than between sentences in the simple past, but the patterns of confusions were very similar. The one deviation noted was that, among simple past sentences, N and PN were more often confused than were K and P, while this did not hold for sentences in the present perfect. Over trials, an increase in the di-

mensionality of the configurations was noted. The constructions tended to become more nearly equidistant, and the configurations could no longer be described perfectly on the basis of the grammatical relationships of the sentences. This change was thought to have been due to a change in the manner by which the sentences were recognized, to a ceiling effect, or to a defect in the transformation used to convert the confusion measures to distances.

Finally, some data presented by Mehler for confusions among sentence constructions in recall were reanalyzed using the Kruskal technique. The resulting configuration was precisely the regular configuration predicted by the Katz and Postal phrase-structure analysis.

Over all studies, much the same relationships were obtained among the sentence constructions investigated. This assertion is supported by the similarity of the scaled configurations, and by the high correlations among the data obtained in Experiments 1 and 2 and by Mehler. In each configuration, constructions analyzed by the Katz and Postal phrase-structure grammar as being more closely related were closer together in the scaled configurations, and the configurations could be dimensionalized in terms of the grammatical relatedness of the sentence constructions. These results indicated a powerful and consistent effect of grammatical relationships among sentences on their perceived similarity.

An explanation of the results on the basis of physical similarity of the sentences was ruled out, primarily on the basis of the comparability of the results obtained with sentences in different tenses and on certain apparent inconsistencies between perceived and physical similarity. However, it was tentatively concluded that physical similarity did have an effect over and above the effect of grammatical similarity. The possibility that the results could be due to the semantic similarity of the sentences rather than to their grammatical relatedness was briefly considered, but it was suggested that no meaningful conclusions could be made at the present time because of the lack of a metric of semantic similarity of sentences.

REFERENCES

ATTNEAVE, F. Dimensions of similarity. *American Journal of Psychology,* 1950, **63**, 516–556.

BACH, E. *Introduction to transformational grammars.* New York: Holt, Rinehart, & Winston, 1964.

CHOMSKY, N. *Syntactic structures.* The Hague: Mouton & Company, 1957.

CHOMSKY, N. On the notion 'rule of grammar'. In R. Jacobson (Ed.), *Structure of language and its mathematical aspects.* (Proceedings 12th symposium in applied mathematics) Providence, R. I.: American Mathematical Society, 1961, 6–24.

CHOMSKY, N. Explanatory models in linguistics. In E. Nagel, P. Suppes, & A. Tarshi (Eds.), *Logic, methodology, and the philosophy of science.* Stanford: Stanford University Press, 1962, 528–550.

CHOMSKY, N. *Aspects of the theory of syntax.* Cambridge: M.I.T. Press, 1965.

CHOMSKY, N., & MILLER, G. A. Introduction to the formal study of natural languages. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2). New York: Wiley, 1963, 269–322.

CLIFTON, C., JR. Syntactic generalization: Affirmation-negation. Unpublished doctoral dissertation, University of Minnesota, 1964.

CLIFTON, C., JR., KURCZ, IDA, & JENKINS, J. J. Grammatical relations as determinants of sentence similarity. *Journal of Verbal Learning and Verbal Behavior,* 1965, **4**, 112–117.

KATZ, J. J., & POSTAL, P. M. *An integrated theory of linguistic descriptions.* Cambridge, Mass.: M.I.T. Press, 1964.

KRUSKAL, J. B. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika,* 1964, **29**, 1–27. (a)

KRUSKAL, J. B. Nonmetric multidimensional scaling: A numerical method. *Psychometrika,* 1964, **29**, 115–129. (b)

LEES, R. B. The grammar of English nominalizations. *Supplement to International Journal of American Linguistics,* 1960, 26.

MEHLER, J. Some effects of grammatical transformations on the recall of English sentences. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2**, 346–351.

MESSICK, S. J., & ABELSON, R. P. The additive constant problem in multidimensional scaling. *Psychometrika,* 1956, **21**, 1–17.

MILLER, G. A. Some psychological studies of grammar. *American Psychologist*, 1962, **17**, 748–762.

MILLER, G. A., & CHOMSKY, N. Finitary models of language users. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology* (Vol. 2). New York: Wiley, 1963, 419–492.

MINK, W. D. Semantic generalization as related to word association. *Psychological Reports*, 1963, **12**, 59–67.

ODOM, PENELOPE. Syntactic generalization: Queries. Unpublished doctoral dissertation, University of Minnesota, 1964.

SHEPARD, R. N. Similarity of stimuli and metric properties of behavioral data. In H. Gulliksen & S. Messick (Eds.), *Psychological scaling: Theory and method*. New York: Wiley, 1960, 33–43.

SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 1962, **27**, 125–140. (a)

SHEPARD, R. N. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 1962, **27**, 210–246. (b)

SHEPARD, R. N. Analysis of proximities as a technique for the study of information processing in man. *Human Factors*, 1963, 33–48.

TORGERSON, W. S. *Theory and methods of scaling*. New York: Wiley, 1958.

(Received June 8, 1965)

Vol. 80, No. 6 Whole No. 614, 1966

Psychological Monographs: General and Applied

EFFECTS ON THE SUBSEQUENT PERFORMANCE OF NEGOTIATORS OF STUDYING ISSUES OR PLANNING STRATEGIES ALONE OR IN GROUPS[1]

BERNARD M. BASS

Graduate School of Business, University of Pittsburgh

In 3 successive experiments, as representatives of management or labor, 256 graduate business students bargained individually with counterparts on 9 issues. 2 of the 4 treatments of each experiment required groups of Ss to plan strategies or to study the issues without considering bargaining tactics. Various kinds of prenegotiation study groups were contrasted. Also, some Ss planned strategies or studied alone rather than in groups. In the 1st and 3rd experiments in which deadlines were imposed, those negotiators who had prepared themselves by planning strategies were more likely to deadlock, more so if they had planned in advance in groups rather than alone. Detailed analyses are presented of the effects of the treatments within each experiment on specific contract outcomes, the overall favorability to the company of the settlements, the departure of the agreements reached from community norms and the speed of settlement. The latter 2 outcomes (departure and speed) were highly correlated. A variety of treatment effects appeared, some of which were consistent across experiments. Also, agreement of each 2 negotiators on the relative importance of issues depended on prenegotiation treatment as did the judged importance of most of the 9 issues and the postsettlement evaluation of the adequacy of the settlement reached. Personal orientation of the negotiators also affected outcomes. Thus, task-oriented negotiating pairs reached settlement closer to community norms while self-oriented negotiating pairs tended to agree more closely on the importance of the issues. Company and union representatives favored using different tactics with different concerns in mind.

SEVERAL rather independent approaches have developed in the study of intergroup conflict resolution. Economists, interested in the exchange of value, have pursued rational and deductive formulations of the problem (Rapaport, 1960; Schelling, 1957; Boulding, 1962) with heavy emphasis on the mathematics of game theory. Political scientists like Mack and Snyder (1957) have extracted generalizations from surveys of historical materials. Psychologists like Sherif et al. (1961) and Blake and Mouton (1961) have focused on the socioemotional aspects of in-group, out-group identification or the implications of reinforcement theory (Osgood, 1962).

Common Elements of Bargaining

Intergroup conflict contains a number of common rational and emotional elements whether it occurs between nations, between union and management, or between department heads of the same company who must negotiate transfer prices for goods that one head is transferring to the other.

The conflicting groups share a common fate. Agreement must be reached if either party is to survive and prosper (Siegel & Fouraker, 1960). Rival nations, competing unions and managements, or contentious department heads must resolve their conflict to avoid mutual social and economic losses.

Typically, bargainers are engaged in a complex non-zero-sum game. Both lose if each seeks to maximize his own gain at the

[1] This work was supported by Contract Nonr 624(14), Group Psychology Branch, Office of Naval Research. Many contributed to various phases of this study. Robert Voytas was particularly helpful in earlier analysis as were Sue Wolpert, Richard Karppinen, and Walter McGhee in later stages. Avi Porat was responsible for interviewing individual negotiators and content analyses of their reported tactics. I am also indebted to my professional colleagues, James Vaughan and Raghu Nath, for their cooperation and aid.

expense of the other. Both parties can profit, although not maximally, by means of a cooperative solution. Both gain when they compromise at less than maximum return for each. Yet, there is no guarantee that the non-zero-sum game will produce cooperating bargainers. On the contrary, competitive strategies are often maintained to the detriment of all concerned (Scodal, Minas, Ratoosh, & Lipetz, 1959). The bargainers will cooperate only if they can develop mutual trust through appropriate communications and if they are oriented toward each other's welfare (Deutsch, 1957; Loomis, 1957).

Organizational commitments. The goals and constraints on our actions often are dictated by organizational considerations. Little variance in behavior is left to personal idiosyncracy when we act as organizational representatives. Surprisingly, this aspect of bargaining has been ignored generally by those primarily interested in the rational elements of bargaining. Nevertheless, bargainers usually negotiate as representatives of their respective organizations. Their bargaining is strongly influenced by their group commitments. The industrial relations director at a collective bargaining session is constrained to a great degree by higher management authority and his management peers with whom he may already have marked out the limits of what he may do. The union representative knows he must strive to achieve a resolution satisfactory to the rank and file (Gouldner, 1954). Negotiators drawn from competing experimental groups in a zero-sum "you win, I lose" game are completely locked into conflicting positions by group identifications. Hardly ever can two representatives agree which of the two groups they represent did the better job, for instance, in preparing an essay on an assigned topic. Each remains committed to his own group's product. As group representatives, subjects are seriously biased in favor of their own group in the evaluation of the situation. Their flexibility is impeded by loyalty to their own group. Deviation from their own group position is treasonous. Their unwillingness to compromise is supported by fear of censure from their own group. Even after studying opposing points of view, these partisans see more divergences than actually exist between their own and other positions. The inability of a negotiator representing a group to agree that his opponent's group did a better job is not necessarily a conscious bias out of fear of sanction by his own members if he were to capitulate, for these biased evaluations will appear to the same degree even if complete secrecy is maintained about the source of the decisions (Blake & Mouton, 1961).

The significance of group commitment to understanding the behavior of the individual negotiator representing the group is emphasized by noting the much lower degree of partisanship exhibited by the individual bargainer who is representing only himself. Thus, Vegas, Frye, and Cassens (1964) showed that two individuals competing as individuals about which one wrote the best essay have relatively little difficulty in agreeing that one or the other paper was best. Here, there is little overevaluation of one's own product and devaluation of the opposing entry. This is in marked contrast to what happens when group representatives discuss the merits of their respective group products. The perceptual distortion which takes place in the evaluations of a negotiator who comes out of a group to represent it in the bargaining process is much less likely to appear when individuals are representing only themselves in negotiations. The inability of negotiators to reach agreement, to perceive issues in the same way, often lies in their group commitments, identifications, and loyalties.

PURPOSE

If negotiators from competing groups were to be freed from these perceptual distortions as well as conscious fears of sanction, it was reasoned that we would need to understand and control the group process which ordinarily precedes negotiations.

To develop such understanding was the purpose of the three successively planned experiments to be reported. The first experiment set out to see if bargainers could avoid the hardening of lines and commitments to partisan positions, if before ne-

gotiations they met together for joint study of the issues dividing them, spending their time in informative discussions rather than in tactical maneuvers or in the planning of such maneuvers. Making both sides study the issues together before beginning to bargain, it was argued, might increase their tendency to focus more on the common interests of both sides and less on difficulties that do not really exist. Also, such bilateral study among future negotiators offered all an opportunity to become acquainted personally with those with whom one would subsequently negotiate.

This first experiment looked at bilateral study among those who would subsequently face each other in negotiations. Three other situations were created and compared with bilateral study in their effects on subsequent negotiations. In one such treatment, joint study was afforded, but not with persons who subsequently would be met in bargaining sessions. This was to see the effect of personal familiarity on the behavior of negotiators independent of the effects of joint study, per se. In another treatment, unilateral study groups were set up so that future negotiators studied the issues but only with members of their own side. A third condition attempted to simulate the ordinary prebargaining strategy meeting of groups facing forthcoming negotiations. (The American Management Association, 1963, advertises a training course for executives in preparing "your strategy for presenting management demands . . . how best to set company goals in your prebargaining sessions. . . .")

When this first experiment found that unilateral study softened conflict as much as did joint study, it suggested the second experiment which examined what happens when one side studies the issues while the other plans strategies. Also, the second experiment was conducted without deadlines to see their importance to differential outcomes. The third experiment returned to the fundamental question of how important was group commitment to a negotiator's attachment to a strategy. It compared negotiators who had studied the decisive issues alone or in unilateral groups before bargaining. It also compared those who planned strategies alone before bargaining with those who planned in groups beforehand.

METHOD

A non-zero-sum union-company bargaining game created by Campbell (1960) was modified and employed to test the differential effects of the treatments in the three experiments. Minor modifications made each experiment somewhat different, so that, generally, statistical comparisons should be limited to treatments within the same experiment.

The Problem

All participants were given a page of background information about the Townsford Company, a small textile firm, and its union, concluding with the paragraph:

> The three year contract has now expired. Negotiations broke down in the final week with both sides adamant in their positions. The only agreement reached was that each side would select a new bargaining agent to represent it, scheduled to meet today (the first day of strike) in an attempt to reach a quick solution and avoid a long strike.

Contract Issues

There were nine issues for bargaining: hospital and medical plan, wages, sliding pay scales to conform to cost of living, seniority, union representative on the Board of Directors, nightshift differential, vacation pay, establishment of a work rules committee, and a checkoff system. Each participant received a graphic statement of the current union and company positions on each issue and the financial cost to the company in thousands of dollars for a 2-year period. For example, for the wages issue, it was as follows:

PAST CONTRACT: $1.94 per hour
UNION: demanded an increase of 16 cents per hour
COMPANY: refused outright

	cents increase per hour									
COMPANY	00	02	04	06	08	10	12	14	16	UNION
Estimated total value in thousands of dollars for two years	(0)	(8)	(16)	(24)	(32)	(40)	(48)	(56)	(64)	

Participants also received data on each of the nine issues for four other local textile plants in the same community and averages for other industries in the same city. Two of the other textile plants employed the same type of workers as the Townsford Company.

Five of the nine issues involved money. Four others, like the question of seniority, did not.

In addition, union representatives received a more detailed, one-page memorandum explaining the union's position, while company representatives received a one-page company memorandum explaining the company position in more detail.

A summary of union and company positions and the normative position defined by the two other plants with the same type of workers are shown in Chart 1.

Subjects

Prior to presenting the bargaining problem, the 256 subjects, all graduate business students, were assigned as the union or company representatives according to whether their scores on a 42-item questionnaire about union-management attitudes developed by Hepler (1953) and refined by Campbell (1960) were above or below the sample median. This was to increase the identification of subjects with the position they had to take as representatives.

As might be expected, the 256 graduate business students as a whole were more pro management ($\bar{X}$ = 117.4) than the 132 undergraduates ($\bar{X}$ = 120.0) drawn from psychology classes by Campbell, but the range of attitudes was about the same for both samples.

Fifty percent of the 256 graduate business students scored between 101 and 126. Ninety percent scored between 92 and 135. Their median score was 114. They tended to pile up cases in the moderate pro management portion of the scale while students from psychology classes tended to concentrate more heavily in the moderate pro union region.

The Assignment

Assembled in a large classroom, the subjects were instructed as follows:

> Your are going to take part in a study of collective bargaining. You will be assigned as a union or company representative depending on your expressed attitudes in the questionnaire you completed two weeks ago.

Participants were given 5 minutes to read the background and contract information described before. Then a copy of the contract itself was given each participant, and the instructions continued.

> In order to settle an issue, both negotiators must accept some specific position on the issue. When both negotiators are in agreement on an issue, one man should read aloud the issue and the position to be endorsed. He then circles on both copies of the contract the position to be endorsed, and each man initials the item in the space provided at the right on both copies. Once both negotiators have agreed and initialed the issue, it is settled for the two-year contract period, and it may not be changed later in the negotiations. Any man may open the discussion and any man may read and circle the position of the issue once agreement is reached. These procedures have been established by joint agreement of the union and company.
>
> *Issues for Bargaining* (Chart 1) is a list of the issues you are to settle and a memorandum concerning the issues. You will be given time to examine this information and to take another look at the Background Information.
>
> There are nine issues to be resolved. The issues are not arranged in any order of importance, and you may discuss them in any order or combination you desire. Under each issue you will find a statement of the specific provisions of the past contract and the positions of the company and the union when negotiations ended last week. Next, you will find a scale that shows at the left the present position of the company and at the right the present position of the union on a given issue. Between these extremes some possible compromises are listed for your convenience. And, finally, beneath the scales in the parentheses, you will find estimates of the amounts of money (in thousands of dollars) that each of the possible agreements directly above would cost or gain for your group in two years.

In the first two experiments, negotiators were reminded each 5 minutes of the amount of time being consumed in negotiation. They were to consider each 5 minutes as being equivalent to 1 full day of negotiation. At the end of each 5-minute interval, a loss accrued of an additional $6,000 to each side in wages or profits. No such loss was involved in Experiment 3.

In the first and third experiments, if no contract was negotiated completely in 70 minutes (or 14 simulated days), negotiations were broken off and the strike continued.

The second experiment was run without deadlines so that the negotiations continued until all contracts were signed. The last contract in Experiment 2 was signed in about 130 minutes as a consequence of some prodding by the experimenter.

TREATMENTS

Experiment 1

A total of 33 contracts were negotiated by the 66 subjects who were involved in the first experiment. For Treatment A, prior to negotiations between single union and single company representatives, two unilateral groups, one of nine pro manage-

Chart 1

ISSUES FOR BARGAINING

1. Hospital and Medical Plan:

Past Contract: Company paid 1/4 of cost, employee paid remaining 3/4
UNION: demanded company pay full cost
COMPANY: refused to pay more than 1/4

	proportion of company payment				
COMPANY	1/4	2/4	3/4	4/4	UNION
Total money value	(0)	(6)	(12)	(18)	

2. Wages:

Past Contract: $1.94 per hour
UNION: demanded an increase of 16 cents per hour
COMPANY: refused outright

	cents increase per hour										
COMPANY	00	02	04	06	08	10	12	14	16	18	UNION
Total money value	(0)	(8)	(16)	(24)	(32)	(40)	(48)	(56)	(64)	(72)	

3. Sliding Pay Scale to Conform to Cost of Living:

Past Contract: pay scale is fixed through the term of the contract
UNION: demanded pay increases in proportion to increases in the cost of living
COMPANY: rejected outright

COMPANY	NO	YES	UNION
Total money value	(0)	(20)	

4. Seniority:

Past Contract: straight plant-wide seniority, workers are laid off on the basis of the number of years with the company
UNION: rejected any changes in the seniority principle
COMPANY: demanded some flexibility in the seniority rule; wants to establish departmental seniority (seniority rule would apply *within* departments only)

COMPANY	YES	NO	UNION
Total money value	(0)	(0)	

5. Union Representative on the Board of Directors:

Past Contract: no union representative on the Board
UNION: demands one union representative be appointed
COMPANY: rejected outright

COMPANY:	NO	YES	UNION
Total money value	(0)	(0)	

6. Night Shift Differential:

Past Contract: an extra 5 cents per hour is paid for night work
UNION: demands a 5 cent increase to 10 cents per hour
COMPANY: rejected

	cents increase per hour						
COMPANY	0	1	2	3	4	5	UNION
Total money value	(0)	(1)	(2)	(3)	(4)	(5)	

Chart 1—Continued

7. Vacation Pay:

Past Contract: 2 weeks paid vacation for all workers with one year service
UNION: wants 3 weeks paid vacation for workers with 10 years of service
COMPANY: rejected

	2 wks. for 1 year service	3 wks. for 20 years service	3 wks. for 15 years service	3 wks. for 10 years service	
COMPANY					UNION
Total money value	(0)	(½)	(2)	(5)	

8. Establishment of a Work Rules Committee:

Past Contract: no work rules committee exists
UNION: rejected establishment of committee
COMPANY: demanded establishment of a work rules committee composed of two company representatives, two union representatives and two efficiency engineers from an industrial consulting firm to study and to be responsible for changes in work rules

COMPANY	YES	NO	UNION
Total money value	(0)	(0)	

9. Check-Off System:

Past Contract: workers pay union dues to union representatives on pay day
UNION: demanded a check-off system whereby the company deducts union dues from the worker's pay for the union
COMPANY: rejected the check-off system

COMPANY	NO	YES	UNION
Total money value	(0)	(0)	

ment men and the other of nine pro union men were told to formulate strategy in preparation for subsequent negotiations, as company or union representatives respectively:

> You should use the 30 minutes to plan your bargaining strategy, to formulate a package of agreements, to prepare for concessions and to decide on items on which you feel as a group each man representing you should stand firm.

For Treatment B, one unilateral company group of eight and one unilateral union group of eight were instructed to study the issues as follows:

> You should use the 30 minutes to learn as a group as much as you can about union and company positions. Rather than formulate any strategies for bargaining, the purpose of this 30 minute study is to promote understanding of the other point of view in comparison to your own, to see the areas of greater and lesser disagreement.

For Treatments C and D, 16 union and 16 company representatives were given the preceding assignment, then divided into four bilateral study groups. Each joint or bilateral group contained four union and four company representatives who were instructed as follows:

> In these study groups of union and company representatives, you should devote the 30 minutes discussion time with learning as much as you can about each others' positions. You should do no negotiating or bargaining during this time. The purpose of the study group is to promote understanding of the other point of view in comparison to your own, to see the areas of greater and lesser disagreement. Negotiation will come later.

Following this, for Treatment C, half of these representatives negotiated as individuals with other single individuals from their *own* bilateral study group, while for Treatment D, the other half had to negotiate with counterparts who had been in a *different* bilateral study group.

Subjects of each group knew only about

their own treatment until the postsession critique.

Some strategic planning actually took place in the unilateral study groups, but it involved setting general guidelines within which its members could remain highly flexible. For example, the union unilateral study group set a goal of $61,500 in monetary concessions to be obtained from the company, but no specific way of attaining this was decided.

Experiment 2

For all four treatments in this second experiment, there was composed a total of 16 six- or seven-man strategy or unilateral study groups, 8 groups representing the union and 8, the company. For each treatment, two groups met for study or planning for each side. A total of 102 subjects who subsequently negotiated 51 contracts were involved in this second experiment.

Treatments A′ and B′ were near-replications of Treatments A and B of Experiment 1. The difference was that in Experiment 2 there were no negotiating deadlines to be taken into account in planning strategies (Treatment A′) or studying the issues in unilateral study groups (Treatment B′). In Treatment E, the two groups of company men planned strategies before bargaining as individual negotiators while the two groups of union representatives studied the issues unilaterally beforehand. Treatment F was the reverse of Treatment E. The two company groups studied the issues beforehand while the two union groups planned strategies in advance of negotiations.

Experiment 3

A total of 88 subjects took part in this third experiment. Of these, 24 underwent Treatment A″, negotiating 12 contracts after having planned strategies as union or company representatives. Twenty negotiated 10 contracts after having studied the issues in unilateral union or company groups of five men each (Treatment B″). The remaining 44 planned strategies alone (Treatment G) or studied the issues alone (Treatment H) in equal numbers of 11 each for each treatment and to provide the necessary union and company representatives for subsequently negotiating 22 contracts.

Inadvertently, no mention was made of the cost of negotiating time in this third experiment. For this reason, Treatments A″ and B″ are not directly comparable with Treatments A′ or A and B′ or B of the preceding experiments as the effect of such announced costs, per se, on negotiations are unknown although they may be quite minimal according to Campbell (1960). He found that when stated costs per simulated day were $3,000, negotiators did not behave significantly differently than when costs were $6,000.

Negotiations

Following the 30-minute study or strategy session, 128 representatives negotiated with 128 opposite representatives, in pairs, until a contract was signed or (in the first and third experiments) the paired negotiators deadlocked at the end of 70 minutes of bargaining. Actually, 7 of the 33 pairs of the first experiment failed to sign contracts in the 70 minutes and 5 of the 44 pairs of the third experiment failed to sign contracts.

A postnegotiations questionnaire was completed by all participants concerning how long they would like to see the contract last, how skilled they felt the other negotiators had been, how acceptable the contract was to one's side, who got the better deal, the defensibility of one's position, and how congruent the role they had played had been with their own beliefs. They also ranked the nine bargaining issues in order of importance.

A postsession critique was conducted with negotiators of a given experiment assembled together and by individual interviews.

Analyses of Variance

Four objective and 15 subjective variables were examined. The time to complete contracts, the absolute departure of the contract settlement from the "going rate," the favorability of the settlement to the company and the extent negotiators agreed on the relative importance of the nine bargaining issues were the objective data gathered.

Two sets of subjective responses were obtained concerning each contract—one from a union negotiator, the other from a company negotiator. These included the following 15 responses:

1–9 Rank in importance assigned to each bargaining issues

10 Length of time in months the respondent would like to see the contract in force

11 Acceptability of contract to one's side (1 = unacceptable; 2 = partially acceptable; 3 = highly acceptable; 4 = fully acceptable)

12 Estimate of who got the better deal in the settlement (1 = union; 2 = both about equal; 3 = company)

13 How easy it was to defend one's position (3 = own position more defensible; 2 = both equal; 1 = own position less defensible)

14 Congruence of assigned role to one's real beliefs (5 = completely; 4 = highly; 3 = somewhat; 2 = slightly; 1 = not at all)

15 Quality of the opponent's performance as a negotiator (1 = very poor; 2 = poor; 3 = fair; 4 = good; 5 = very good)

For each of these variables in Experiment 1, analysis of variance contrasted the effects of the four treatments. The degrees of freedom were distributed as follows when data were available for all 33 contracts or 66 negotiators.

Objective variable	df
(i) Between treatments	3
Error	29
Total	32

Subjective variable	df
(i) Between treatments	3
(j) Union or company respondent	1
i × j	3
Error	58
Total	65

In Experiment 2, the analysis of objective data was as follows for 48 of the 51 contracts. To provide an equal number of cases per treatment (12), a total of three contracts were withdrawn randomly from the subjects of Treatments B′ and F.

Objective variable	df
(i) Union strategies or studies	1
(j) Company strategies or studies	1
i × j	1
Error	44
Total	47

Subjective variable	df
(i) Union strategies or studies	1
(j) Company strategies or studies	1
(k) Union or company respondent	1
i × j	1
i × k	1
j × k	1
i × j × k	1
Error	88
Total	95

For Experiment 3, the objective analysis was as follows for 36 contracts of the 44 from which 9 were withdrawn because of deadlocks or at random to equalize the number of cases in each treatment:

Objective variable	df
(i) Strategies or studies	1
(j) Alone or in groups	1
i × j	1
Error	32
Total	35

Subjective variable	df
(i) Strategies or studies	1
(j) Alone or in groups	1
(k) Union or company respondent	1
i × j	1
i × k	1
j × k	1
i × j × k	1
Error	64
Total	71

Results

Mean differences among the experimental treatments will be discussed in order in the text that follows. The relevant completed analyses of variance for Experiments 2 and 3 are tabulated for reference as an Appendix. Such complete analyses were not undertaken for Experiment 1, and only the results of such analyses that were done for

Experiment 1 will be stated in the text. The tabled analyses in the Appendix are as follows:

Table: Experiment 2	Table: Experiment 3	Analyses of Variance of the Effects of Experimental Treatments on:
A1	A5	Contract outcomes
A2	A6	Specific settlements
A3	A7	Satisfaction with outcomes
A4	A8	Ranked importance of issues

Objective Results

Speed of Conflict Resolution

It was assumed that for comparable conditions, the firmness of two negotiators' commitments to their respective groups or strategies would be reflected in the total amount of time they required to settle the nine issues and to sign the contract. The 12 deadlocks, where no resolution was achieved, necessitated calculating harmonic mean times rather than mean times for all cases in order to take into account the absence of values for the deadlocked cases. (The harmonic mean is the reciprocal of the mean of the reciprocals of the original values. Deadlocks are assumed to be settled in an infinite amount of time. The reciprocal of infinity is zero, so the deadlocks now can be treated as zero values and included in the distribution with the reciprocals of all other obtained values. The harmonic mean time can be interpreted as the *speed* with which the conflict was resolved where the mean time is the length of time required to resolve the conflict. Where many deadlocks occurred as in Treatment A of the first experiment, a highly inflated harmonic mean was produced as seen in Table 1.)

Where deadlines were employed in the first and third experiments, it can be seen from Table 1, they produced deadlocks. These deadlocks did not appear haphazardly. They never occurred for Treatments G and H where individuals planned or studied alone prior to negotiations. They occurred most often when negotiators had been in strategy planning groups before negotiating as individuals.

When no deadlines were applied as in Experiment 2, the differential effects disappeared on speed of negotiation of Treatments A′ and B′, group strategy planning versus group study. The obtained differences in harmonic mean times between Treatments A and B and A″ and B″ of the first and third experiments hinged on the differential amounts of deadlocks generated by the respective treatments. Subsequent critiques of the strategies formulated by opponents suggested that strategy planning led to relatively less flexible positions. If deadlines were imposed and *if the strategists had nonoverlapping strategies, deadlocks* were likely. On the other hand, if the opposing strategists had highly overlapping strategies, very speedy resolution was possible. For example, if the company's strategy was to hold firm on no wage increase and the union's strategy was not to settle for less than a 6-cent increase, a deadlock was probable. Conversely, if the company strategy was to yield a maximum of an 8-cent increase on wages and the union was happy to settle for 6 cents, a more rapid negotiation of this issue was likely than if opponents had no fixed plans and had only studied the issues. In short, the effects of strategic thinking could produce deadlocks on the one hand, or faster-than-average resolutions on the other. Where deadlines were imposed, the overall effect on the calculated harmonic means was to yield slower speeds of negotiating for strategists than for those who studied the issues. Where no deadlines were imposed (Experiment 2), strategic thinking led to as fast or faster resolution than studying the issues—although, to repeat, this mean difference depended on the contents of the opposing strategies involved.

In all three experiments, harmonic mean times varied significantly at the 1% or 5% level as a function of treatment. However, in Experiment 1, the three methods of studying the issues did not yield differentially significant speeds of settlement. In Experiment 2, fastest resolution occurred when company strategists faced union representatives who had studied the issues. In Experiment 3, the greater source of variance was associated with whether or not the negotiators had been alone or in

TABLE 1

EFFECTS OF PRIOR EXPERIENCE OF NEGOTIATORS ON CONTRACTS THEY NEGOTIATE SUBSEQUENTLY

	(1) Number of contracts negotiated	(2) Deadlocks	(3) Harmonic mean time in minutes to negotiate contracts	(4) Absolute departure of settlement from going rate	(5) Favorability to company of settlement excluding cost of negotiations[a]	(6) Agreement on relative importance of issues
Experiment 1 (all in groups; deadlines)						
A Plan strategy	9	5	163.0	230.5	49.0	.60
B Unilateral study	8	1	24.5	171.9	78.2	.52
C Bilateral study (with future opponents)	8	0	30.5	185.5	64.5	.76
D Bilateral study (with others)	8	1	26.5	170.1	79.9	.38
All	33	7	61.0	183.5	67.9	.61
Experiment 2 (all in groups; no deadlines)						
A′ Plan strategy	12	—	38.0	264.1	116.4	.50
B′ Unilateral study	14	—	41.3	245.1	40.2	.46
E Companies plan strategy, unions study	12	—	29.6	262.2	125.7	.24
F Companies study, unions plan strategy	13	—	38.7	295.2	57.7	.36
All	51		37.1	266.6	85.2	.39
Experiment 3 (deadlines: no cost associated with negotiation time)						
A″ Plan strategy (in groups)	12	3	56.2	265.3	0.4	.54
B″ Unilateral study (in groups)	10	2	51.9	288.0	91.8	.48
G Plan strategy (alone)	11	0	44.8	235.2	63.7	.40
H Study (alone)	11	0	36.3	337.2	125.0	.41
All	44	5	47.4	281.5	70.7	.46
Grand totals	128	12	46.8	250.3	75.8	.47

[a] Minus signs have been omitted from in front of all values in this column to make reading easier. To interpret, the higher the value in the column, the more favorable the settlement to the company.

groups rather than whether they had planned strategies or studied. All five deadlocks occurred for negotiators with prenegotiation experience in groups. No deadlocks occurred for those who had worked alone prior to negotiating. There was a cumulative effect of treatments as can be seen in Table 2. Group strategists took longest; individuals who studied alone were fastest in negotiations.

In all three experiments there was a subjective indication of greater conflict between those who had planned strategy than between those who had studied the issues. Although no objective measures were made, the experimenter and assistants sensed that the noise level generated by the arguing of negotiators from strategy groups was far more intense than the noise created by the arguments of the negotiators from study groups. In sum, in comparison to those who only studied the issues, negotiators committed to strategies could negotiate as rapidly, particularly if their strategy overlapped their opponent's strategy, but they were more likely to be caught in deadlocks if forced to negotiate against deadlines.

Direction of Settlement: Calculations

A scoring procedure developed by Campbell (1960) was applied to each of the 116

TABLE 2

Harmonic Mean Time, in Minutes, to Reach Conflict Resolution by Negotiators as a Function of Previous Experience in Studying Issues or Planning Strategies Alone or in Unilateral Groups (Experiment 3)

	Prenegotiation Experience		
	Group (N = 21 contracts)	Alone (N = 23 contracts)	Both (N = 44 contracts)
Study (N = 22 contracts)	51.9	36.3	42.7
Strategy (N = 22 contracts)	56.2	44.8	49.3
Both (N = 44 contracts)	53.8	40.3	45.8

signed contracts. One score indicated how much the agreement reached by the two negotiators on each issue deviated *algebraically* from the "going rate" in the other two comparable local textile plants in the same community. Another score indicated the *absolute* deviation of the settlement from the "going rate."

The deviation values were obtained by assigning zero to a settlement of an issue at the going rate and converting succeeding intervals departing from this zero point into percentages of 100 units. If there were four deviating points on a scale for an issue, then the deviations from the going rate were 25, 50, 75, and 100; for five points, they were 20, 40, 60, 80, and 100. If the going rate, or zero deviation, was between what the union and management were demanding, then the union position arbitrarily was positive (more than the going rate) and the management, negative (less than the going rate). Therefore, the lower the sum of algebraic values of deviations for the nine contract issues from the going rates, the more the settlement favored management; the higher the sum of algebraic deviations, the more the agreement favored the union. When signs were taken into account, all mean settlements were negative, at less than the going rate, that is, favorable to the company. Column 5 of Table 1 shows the obtained means as a function of the treatments of each experiment omitting the minus signs in front of each of the means. As displayed in Table 1, the higher the numbers the better the outcome for the company.

When the signs were ignored in calculating the sum of the absolute deviations for each contract, the value indicated simply how much the negotiators departed from the going rate, in one direction or the other, in settling each issue in order to reach a final agreement. Column 4 of Table 1 shows the aggregate results. They exclude, of course, the 12 deadlocked sets of negotiations since no contracts were agreed to by the deadlocked parties.

While treatments within experiments were a significant source of variance in the departure of the contracts from the going rate, similar treatments failed to maintain their differential effects in the three experiments.

Departure from the Going Rate: Results

As seen in Table 1, Column 4, for the first experiment the four bargaining pairs from the strategy groups who reached agreement had to depart more widely from the going rate than did the bargainers from study groups, as a whole. The mean absolute deviation for these four pairs of strategists was 230.5 while it was 175.8 for all those from study groups ($p < .05$).

Again, in the second experiment, those negotiators from strategy planning (Treatment A′) reached agreements departing more (264.1) from the going rate than those who came from Treatment B′ unilateral study groups (245.1), but these results failed to attain statistical significance, and, in the third experiment, results were opposite. Those who studied (Treatments B″ and H″) reached agreements scoring

on the average at 312.7. This was statistically greater at the 5% level of confidence than the value of 250.3 attained for those who underwent the strategy Treatments A″ and G.

Significantly greater departures, 266.6 and 281.5, occurred for Experiments 2 and 3 than for Experiment 1 where the mean value was 183.5. It may be that negotiators were more prone to reach settlements closer to normative solutions to the various issues when they had combined pressure on them to settle quickly, that is, deadlines and announced costs for each 5 minutes they used to negotiate (Experiment 1). When they could operate without deadlines (Experiment 2) or they could use time to negotiate without any stated cost (Experiment 3), they seemed more willing to accept resolutions departing more widely from going rates.

Favorability of Settlement to Company (Column 5, Table 1)*: Results*

When deadlines were in operation (Experiments 1 and 3) settlements favoring the company were significantly ($p < .05$) more likely to be reached when negotiators had studied the issues beforehand rather than planned strategies. In Experiment 1, settlements averaged 74.2 for the 24 contracts between parties who had studied the issues and 49.0 for the four settled contracts between strategists. In Experiment 3, those who studied the issues (alone or in groups) yielded an average score of 108.4 in favor of the company, which average was significantly greater ($p < .05$) than the 32.1 obtained for those who planned strategies before negotiating.

Nevertheless, a complete and significant reversal occurred in Experiment 2, where there were no deadlines. Solutions favoring the company were most likely ($p < .01$) if the company representatives planned strategies (121.0), even more so when the company but not the union men planned strategies (125.7). Resolutions favored the union if both sides studied (40.2) or if the union planned strategies while company representatives studied beforehand (57.7).

The grand mean for all 116 settlements of 75.8 was clearly in the direction favoring the company as were these values for all 12 treatments. This outcome may be an inherent game characteristic or a consequence of the modal attitude towards unions and management of the sample of 256 graduate business students who served as negotiators for the three experiments.

Specific Outcomes

It was possible to do a more detailed analysis of the completed contracts of Experiments 2 and 3. Grand mean settlements of each issue and significant effects were as follows:

Hospital and medical plan. The union had demanded that the company pay full cost; the company had refused to pay more than 25%. The mean settlement reached was identical in both Experiments 2 and 3. On the average, the company had to pay 60.5% of the cost. In Experiment 2, when union representatives studied the issue rather than planned strategies, the settlement was significantly ($p < .05$) more in their favor (65.7% versus 55.2% to be paid by the company).

Wages. The union demanded an increase of 16 cents per hour; the company had refused outright. The mean settlements called for a 7.4-cent increase in Experiment 2 and a 6.2-cent increase in Experiment 3. There were no significant treatment effects.

Sliding pay to conform to cost of living. The union had demanded that the scale shift with a rise in the cost of living; the company had refused outright. Eighteen of 48 contract settlements contained sliding scales in Experiment 2, while 22 of 36 contained such scales in Experiment 3. There were no treatment effects.

Seniority. The union had demanded continuance of plantwide seniority; the company proposed departmental seniority. Thirty of 48 contracts maintained plantwide seniority in Experiment 2, while 22 of 36 did so in Experiment 3. Here negotiators from a group that strategized before bargaining were most likely ($p < .05$) to maintain plantwide seniority (8 out of 9) while negotiators from a group that studied the issues were least likely (2 out of 9).

If alone before negotiating, strategy or study made no difference. In both of these cases, 6 of the 9 contracts maintained plantwide seniority.

Union representative on board of directors. The union had proposed this, and the company had refused. None of the 48 settlements of Experiment 2 contained acceptance of this proposal and only 4 of the 36 contracts of Experiment 3 did so.

Night-shift differential. The union had demanded an extra 5 cents for night work; the company had refused. The mean settlement in Experiment 3 was 2.8 cents. There were no treatment effects in Experiment 3. In Experiment 2, the mean was 3.4 cents. Here the best settlement for the union at 4.4 cents occurred when both sides had studied the issues ($p < .01$). The next best settlement occurred when both sides had planned strategies (3.8 cents). Company study in general led to higher settlements averaging 3.9 cents in comparison to a mean of 2.9 cents given up by company strategists ($p < .05$).

Vacation pay. The union had wanted 3 weeks vacation for 10 years service, requiring an increase of $5,000 annually in labor costs. The company had favored the status quo of 2 weeks with 1 year service, requiring no increase in labor costs. The mean settlement cost the company in Experiment 2 an increase of approximately $1,220 and a corresponding $1,130 in Experiment 3. There were no treatment effects.

Work rules committee. The union had rejected establishment of a committee demanded by the company. This committee of company, union, and industrial engineering consultants was to be responsible for changes in work rules. In Experiment 2, 20 of the 48 contracts called for establishing the committee, but 15 of these were contracts negotiated by company strategists and only 5 were contracts negotiated by company men who had studied the issues ($p < .01$). Experiment 3 was a complete reversal. Twenty-five of 36 contracts called for such a committee. But of these, 17 decisions favoring the committee occurred with the 18 contracts among negotiators who had studied the issues beforehand alone or in groups while only 8 of 18 came from those who had strategized before negotiating ($p < .01$).

Checkoff system. The union had wanted the company to deduct dues from workers' pay; the company had refused. In Experiment 2, 27 of the 48 settlements provided for introducing a checkoff system. Twenty-two of 36 agreed to the system in Experiment 3, most often (8 of 9) when negotiators had come from study groups; least often (3 of 9) when negotiators had studied alone. This compared to the acceptance rate of 6 or 7 of 9 by negotiators who had planned strategies beforehand in groups or alone. (The interaction was significant at the 5% level.)

In sum, when treatment did affect specific outcomes, the effects were likely to be complex. Particularly sensitive to treatment effects were the settlements dealing with plantwide seniority, the night-shift differential, the work rules committee, and the checkoff system.

Agreement on Importance of Issues

For each of the 128 pairs of bargainers, the correlations in their rankings of the importance of the nine issues were calculated, then converted to Fisher's Z. The mean correlations for each of the 12 treatments are shown in Table 1, Column 6.

Dealing with attitudes of bargaining pairs toward the nine contractual issues, these correlations probably are less affected by the presence or absence of deadlines, or the presence or absence of negotiating time costs. The correlations were more dependent on how and when the attitudes were shaped. As can be seen in Table 1, if both parties had been together in bilateral study groups, they were in the highest degree of agreement (.76) about the relative importance of the nine issues. Conversely, if one party to a negotiating meeting studied while the other planned strategies as in Treatments E and F of Experiment 2, least agreement was likely (.24 and .36). Finally, in all three experiments, those who planned strategies in groups (Treatments A, A′, and A″) were slightly more in agreement (.60, .50, .54) than those pairs of

negotiators who came from Treatments B, B′, and B″ unilateral study groups (.52, .46, .48). To sum up, agreement on the relative importance of issues by both parties is enhanced if the parties study the issues together beforehand. Less agreement occurs if each side plans strategies, and still less agreement is likely if each side studies the issues unilaterally.

Evidently, while bilateral study fails to produce faster resolution of conflict in comparison to unilateral study, it does seem to commit its specific members from both sides to agreement about the order of importance of issues.

Each bilateral study group emerges with different schedules of commitment of its members. If negotiators come from the same bilateral study group, they are most likely to agree on what issues are important; if they come from different study groups, they have strong commitments to different orders and are most likely to disagree.

Interrelations among Contract Outcomes

Table 3 displays the product-moment correlations among the four contract outcomes just discussed: harmonic mean negotiating time, departure of settlement from going rate, favorability of settlement to the company, and agreement of the opposing parties on the relative importance of the issues.

Certain consistent relations are apparent. The longer the parties spent in negotiating, the less likely were the contract settlements to depart absolutely from the going rates. Correlations were negative for all 12 treatments. If closeness of the settlement to the going rate is regarded as a measure of the quality of the settlement, then it is clear that speedy settlements were incompatible with the quality of the settlements.

In Experiment 1, fast resolution favored the company (−.34), but a scattered pattern of results were obtained for Experiments 2 and 3, making any overall

TABLE 3

INTERCORRELATIONS FOR EACH TREATMENT AMONG DURATION OF STRIKE, DEPARTURE OF SETTLEMENT FROM NORMS, AND AGREEMENT ON ISSUES FOR 116 SETTLEMENTS

	Number of contracts settled	Negotiating time versus departure of settlement	Negotiating time versus favorability to company	Negotiating time versus agreement on issues	Departure of settlement versus favorability to company	Departure of settlement versus agreement on issues	Favorability to company versus agreement on issues
Experiment 1							
A Plan strategy	4	−.99	−.20	−.76	.30	.76	.31
B Unilateral study	7	−.04	−.09	−.32	.04	.22	−.28
C Bilateral study (with future opponents)	8	−.06	−.58	.24	−.36	−.08	.21
D Bilateral study (with others)	7	−.73	−.36	−.40	.22	.12	−.06
All	26	−.54	−.34	−.35	.05	.30	−.05
Experiment 2							
A′ Plan strategy	12	−.44	.12	−.02	−.14	.23	−.36
B′ Unilateral study	14	−.01	.18	−.48	.55	−.32	.01
E Companies plan strategy, unions study	12	−.23	−.72	.08	.66	.20	.15
F Companies study, unions plan strategy	13	−.05	.13	−.44	.43	.30	−.25
All	51	−.19	−.12	−.32	.40	.19	−.12
Experiment 3							
A″ Plan strategy (in groups)	9	−.02	.23	.06	.64	−.40	−.20
B″ Unilateral study (in groups)	8	−.27	.39	−.44	.00	.09	−.02
G Plan strategy (alone)	11	−.30	−.20	.01	.61	−.16	.01
H Study (alone)	11	−.21	−.03	.09	.04	.17	−.34
All	39	−.20	.10	−.08	.38	−.08	−.14

generalizations impossible about negotiating time and favorability of outcomes to the company.

It would stand to reason that the *more* opposing sides agreed on the relative importance of the nine conflicting issues, the *less* time they would need to reach agreement on how to resolve the conflicts. (For instance, it might make it easier for them to formulate and accept package deals.) The expected negative correlations were obtained in seven instances; correlations close to zero (.08, .06, .01, and .09) were obtained for four treatments and only one sizable positive correlation (.24) appeared. The aggregate mean correlations for the three experiments were −.35, −.32, and −.08, respectively. It is inferred that agreement on the importance of issues was associated with speedy negotiations.

Since all settlements tended to favor the company, it again seems reasonable for greater departures from the going rates to have coincided with better settlements from the company's standpoint. For the three experiments, the mean correlations were .05, .40, and .38. However, their inconsistent scatter made interpretation difficult. Similarly, a haphazard pattern of correlations appeared between agreement on issues and departure of settlement and agreement on issues and favorability of outcome to the company. In sum, settlements closer to the going rate were more likely to be achieved if negotiations went more slowly. And such slower decisions were likely when negotiators initially were in less agreement about the importance of the issues.

Subjective Results

Ranked Importance of Issues

As Table 4 shows, there was complete agreement as a whole by respondents from all three experiments on the order in importance of the first three issues and almost as much agreement on the ordering of the last two issues. Wages, sliding pay, and the hospital-medical plan were most important; the union representative on the company board and the checkoff were least.

A question can be raised about the applicability of the analysis of variance model

TABLE 4

Ranked Importance of the Nine Bargaining Issues in the Three Experiments

Issue	Experiment		
	1	2	3
Wages	1.52	1.72	1.49
Sliding pay scale	2.52	3.53	2.96
Hospital-medical plan	4.00	3.85	4.22
Seniority	5.42	4.56	4.85
Work rules committee	6.31	5.30	5.11
Night-shift differential	4.88	5.61	5.95
Vacation pay	6.19	6.20	6.13
Union representative on company board	6.52	6.14	6.41
Checkoff	7.64	7.94	7.82

since data contributing to the means were influenced by grouping effects assuming that individual respondents were influenced by group discussions prior to negotiating and by common strategies to which they committed themselves. If only one group was represented in a cell, this would reduce the within-cell variance; if several groups were represented, it would probably serve to inflate the within-cell variance.

Treatments had complex effects on the relative importance negotiators attached to some of the various issues. It is important to keep in mind the built-in negative relations among the rankings of the nine issues. If one issue was ranked high, some other issue was forced to be ranked lower.

Wages. In all three experiments this was the most important issue. For the three experiments, average ranks of 1.52, 1.72, and 1.48 were assigned to wages. In the first experiment, the assignment was unaffected by treatment or position as union or company representative.

In Experiment 2, an interesting interaction significant at the 5% level appeared. Wages were judged relatively less important by both union and company respondents whenever they had similar prenegotiation experience and were judged more important when they had different prenegotiation experience. Table 5 shows the outcomes.

In Experiment 3, two significant inter-

TABLE 5

RANKED IMPORTANCE OF VARIOUS ISSUES AS A FUNCTION OF PRENEGOTIATION EXPERIENCE IN EXPERIMENT 2

	48 union respondents		48 company respondents		96 respondents	
	Company plans strategy	Company studies	Company plans strategy	Company studies	Company plans strategy	Company studies
Wages						
Union plans strategy	2.08	1.42	1.83	1.75	1.96	1.58
Union studies	1.42	2.17	1.25	1.83	1.33	2.00
Sliding pay scale						
Union plans strategy	2.83	4.33	3.50	3.00	3.17	3.67
Union studies	6.17	3.33	3.33	1.75	4.75	2.54
Seniority						
Union plans strategy	5.33	3.50	5.83	3.33	5.58	3.42
Union studies	2.83	4.67	5.67	5.33	4.25	5.00
Work rules committee						
Union plans strategy	4.92	5.75	4.08	6.00	4.50	5.88
Union studies	5.00	4.92	4.75	7.00	4.88	5.96
Night-shift differential						
Union plans strategy	6.83	4.75	6.42	4.92	6.63	4.83
Union studies	4.08	6.92	5.00	6.00	4.54	6.46
Vacation pay						
Union plans strategy	7.42	6.00	6.75	6.33	7.08	6.17
Union studies	5.75	5.75	5.17	6.42	5.46	6.08
Union representative on board						
Union plans strategy	2.50	7.67	4.42	8.33	3.46	8.00
Union studies	8.17	4.33	8.75	4.92	8.46	4.63

actions emerged. As shown in Table 6, those who studied in groups or planned strategies alone ranked wages as significantly ($p < .01$) more important than those who studied alone or planned strategies in groups. Above and beyond this, company respondents who worked in groups prior to negotiations and union respondents who worked alone regarded wages as more important.

TABLE 6

RANKED IMPORTANCE OF WAGES BY UNION AND COMPANY RESPONDENTS AS A FUNCTION OF THEIR PRENEGOTIATION EXPERIENCE IN GROUPS OR ALONE

	Group	Alone	Both treatments
Planned strategy	1.60	1.20	1.40
Studied	1.20	1.95	1.58
Company respondents only	1.25	1.95	1.60
Union respondents only	1.55	1.30	1.42

Sliding pay scale. This was the next most important issue in all three experiments. Mean ranks assigned were 2.52, 3.53, and 2.96. While treatments had no significant effects in the first and last experiments, in Experiment 2 company respondents judged the issue significantly more important ($p < .01$) than did union respondents (2.90 versus 4.17), particularly if the company men had studied the issues rather than planned strategies. Table 5 shows the effects of treatments on judgments here.

The triple interaction for Experiment 2 was significant at the 5% level. Thus, the issue was judged most important (1.75) by company respondents who studied and negotiated with opponents who studied; the issue was seen as much less important (61.17) by union men who had studied but who bargained with company strategists.

Hospital and medical plan. For the three experiments, this issue was next most important, ranking 4.00, 3.85, and 4.22, respectively. This time there were no significant effects associated with the first two experiments, but in Experiment 3 union men clearly found this issue more important ($p < .01$) than did company respondents (3.27 versus 5.17). Those who had worked in groups judged the issue more important ($p < .05$) than those who worked alone (3.82 versus 4.62). Those who had studied in groups seemed particularly concerned about this issue ($p < .01$) compared to those treated in all other ways.

Seniority. This was judged fourth in importance in Experiments 2 and 3 and fifth in Experiment 1. Those who worked alone in Experiment 3 prior to negotiating saw this issue as more important ($p < .01$) than those who worked in groups (4.58 versus 5.12). In Experiment 2, company respondents felt more strongly about the matter ($p < .05$) than union men (4.08 versus 5.04), but the most sizable effect seemed a consequence of each side who bargained having had a different prenegotiating experience. As seen in Table 5 for the 96 respondents as a whole, where they both had studied or both had planned strategies before negotiations, they saw the issue as less important ($p < .01$).

Work rules committee. This was judged fifth in importance in the last two experiments and seventh in Experiment 1. In Experiment 2, as can be seen in Table 5, the issue was more important particularly for company men ($p < .05$) where they had planned strategies rather than studied.

Night-shift differential. Judged sixth in Experiment 2 and 3 and fourth in Experiment 1, the union respondents regarded this differential as more important ($p < .01$) than company respondents in both Experiments 1 and 3 but no differently in Experiment 2. For the three experiments, the figures for union and company respondents were, respectively: 4.41 versus 5.34, 5.65 versus 5.59, and 5.25 versus 6.65. Again, a complex interaction was significant ($p < .01$) for Experiment 2 as a function of treatments as Table 5 shows. The nig' '- shift differential was more important o negotiators whose prenegotiating experience had been different rather than similar.

Vacation pay. Judged sixth, seventh, or eighth in importance, vacation pay was seen as more critical ($p < .05$) by all respondents when union men studied rather than planned strategies before negotiating (5.77 versus 6.63) but particularly ($p < .05$) when they had to deal with company strategists. Here, the mean judgment of all involved respondents was 5.46.

Union representative on company board. This was the only issue in Experiment 1 whose importance was significantly affected by treatment ($p < .01$). The proposal that a union representative sit on the company board was more important for strategy planners than those who studied the issues (5.94 versus 6.67). Evidently, this must have been an item pushed by the union during its strategy planning which raised its saliency to union strategists and the company strategists who had to bargain with them relative to its assigned importance by study-group bargainers.

Table 5 shows the highly significant interaction ($p < .01$) appearing in Experiment 2. The issue was seen as far more important for competing strategists (3.46) and for competing negotiators, both of whom had studied the issues (4.63), than for those who had been treated differently (8.46 and 8.00). In Experiment 3, the large effect ($p < .01$) was evidenced in the greater regard shown for the issue by company than union respondents (5.10 versus 7.72).

Checkoff. This was judged least important in all three experiments but effects were mixed in each experiment. The union attached significantly ($p < .01$) more importance to it (7.13) than did the company which ranked it 8.16 in Experiment 1. In Experiment 2, the reverse occurred with company respondents ranking the issue at 7.61 and union respondents ranking it 8.27. In Experiment 3, no significant differences emerged in this response.

In general, result judgments about the mportance of the issues were similar in all three experiments and were relatively unaffected by differential treatments. Wages, sliding pay scales and the medical plan were judged more important, while the proposals

for checkoff systems, union representation on the board and vacation pay were judged least important.

Preferred Length of Time for the Contract to Run

In the first and third experiments, regardless of treatments, company respondents favored significantly ($p < .01$) longer-running contracts than did union men: 36.6 and 33.2 months for company men versus 22.8 and 21.8 months respectively for union men. However, the same difference failed to appear in the second experiment. Here, a significant interaction effect materialized ($p < .01$) for respondents as a whole. Shorter contracts of 24.5 and 22.8 months were favored when each side had had a different prenegotiation experience: union study-company strategy and union strategy-company study. When both had planned strategies or when both had studied, they favored longer contracts (39.0 and 37.5 months respectively).

Acceptability of Contract Settlement to One's Own Side

How confident a negotiator was that his settlement was acceptable to his own side was significantly greater ($p < .05$) if he had been involved in Experiment 2 with negotiations with a company strategist or as a company strategist rather than with or as a company man who had studied the issues (2.83 versus 2.54). No other significant effects were observed in any of the three experiments.

Estimate of Who Fared Better in the Settlement

In Experiment 2, contrary to objective outcomes, company respondents felt the union came out best when the union had studied the issues (1.71), while the company got the better deal when the union had planned strategies (2.09). Union respondents felt the opposite reporting that the union fared best when it planned strategies (1.83) and not as well when it studied the issues (1.92). This interaction was significant at the 5% level.

In Experiment 3, company respondents felt that when they had worked alone before negotiating, the company got the best deal (2.35); when they studied or planned strategies beforehand in groups, the union came out best (1.75). Union respondents showed no significant differences as a function of treatment from a union grand mean of 2.02. Moreover, this mean indicated that they felt a sense of equity in the outcome. Again, the overall interaction effect was significant at the 5% level.

Defensibility of One's Own Position

In Experiment 2, the higher order interaction was significant at the 5% level. Union men felt their assigned position was least defensible (1.33) when they and their opponents had planned strategies beforehand and most defensible when they had studied the issues while the company planned strategies (2.42). On the other hand, company respondents felt least defensible when both they and their opponents had studied the issues (1.83) and most defensible when they had studied and the unions strategized (2.33).

In Experiment 3, union men as a whole felt themselves to be in a significantly more defensible position ($p < .05$) than company respondents felt (2.30 versus 1.95). But beyond this, among both, men from prenegotiation group experience felt themselves in significantly ($p < .05$) more defensible positions (2.05) than those who had studied or planned strategies alone beforehand (1.70). (Group support evidently increased confidence in one's own position.)

Congruence of Assigned Role to Own Beliefs

The company rather than the union role was more compatible in all three experiments with the beliefs of these business school subjects, but the differences in compatibility were significant (and to a modest degree, $p < .05$) only in Experiment 3. In Experiment 2, the biggest effect ($p < .01$) was associated with the interaction of prenegotiation treatments. Congruence was higher (2.83 and 2.96) when union role players studied and company role players planned strategies and vice versa. Con-

gruence was lower (1.83 and 2.17) when both sides planned strategies or studied beforehand.

Quality of Opponent's Performance as a Negotiator

Union strategists of Experiment 2 judged their opponent's performance as significantly ($p < .01$) better (4.33) than did union men who had studied beforehand (4.02). In Experiment 3, opponents were judged more favorably by union men who had worked alone (4.6) and by company strategists who had worked alone (4.7). They were judged significantly ($p < .05$) less favorably by union men with group experience (4.10).

Orientation of the Bargainers

Self-, interaction-, and task-orientation scores (Bass, 1962) were available for the negotiators of Experiments 1 and 3. To see the extent orientation of the bargaining team affected contract outcomes, the sums of self-orientation scores for each bargaining pair and the differences between the self-orientation scores for each bargaining pair were calculated. The same was done for the interaction- and the task-orientation scores. Then, for each treatment, these six values were intercorrelated with contract outcomes.

Again, consistent patterns were sought since the likely significance of a single correlation based on eight cases is quite low.

Task Orientation and Negotiating Behavior

For the eight treatments, A, B, C, D, A″, B″, G, and H, the correlations between the combined task-orientation of the pairs of negotiators and the departure of the settlement reached from the going rate were respectively: −.49, −.28, −.55, −.71, .06, .34, −.44, and −.49. On the other hand, the *differences* in task orientation of the pairs correlated respectively: .27, .14, .28, .78, .44, −.11, .17, and −.35 with this departure. Thus, task-oriented negotiators tended to reach settlements closer to the going rate. (Interestingly enough, Campbell (1960) regards closeness to the going rate as a criterion of the quality of the resolution.) But the two negotiators were less likely to reach settlements closer to the going rate, if they differed widely from each other in task-orientation scores.

At the same time, consistent with the meaning of task orientation, under most treatments, pairs of task-oriented negotiators with high combined task-orientation scores tended to regard vacation pay as less important by assigning the issue a lower rank, that is, 9, 8, 7, 6 rather than 1, 2, 3, or 4. Therefore, the generally positive correlations obtained of .20, .21, .55, .35, .14, .00, .80, and .00 showed high task-orientation among negotiators to coincide with the assignment of low importance to the issue.

Self-Orientation and Negotiating Behavior

When the negotiating pair was high in self-orientation according to their combined scores, they appear to agree slightly more on the importance of the nine issues. The correlations between combined self-orientation and the agreement of pairs of negotiators for the eight treatments were, respectively, .22, .46, .18, .28, .41, .11, −.08, and .26.

When combined self-orientation was high among negotiators, they also tended to attach less importance to the work rules committee. Correlations for the eight treatments were: .52, .48, .30, −.56, .00, .23, .20, and .76. Likewise, they attached less importance to the checkoff system; correlations for the eight treatments were, respectively, .13, .12, .21, .12, .07, .62, .12, and .46.

Consistent with their self-concerns, under seven of eight treatments, self-oriented negotiators ranked the hospital and medical plan as more important an issue. Correlations were −.15, −.21, −.13, −.18, −.67, −.17, −.20, and .39 between the rank assigned the issue and combined self-orientation scores.

Self-oriented negotiators seemed to show different patterns as a consequence of whether they were from a strategy planning group or not. In Experiment 1, strategy pairs with high self-orientation scores reached settlements more favorable to the company (.32) at lower costs to the company (.47), but self-oriented negotiators from study groups did the reverse. For them, correlations between self-orientation com-

bined scores correlated −.33, −.69, and −.40 with degree of favorableness of the settlement for the company, and the combined scores correlated −.31, −.74, and −.18 with costs of the settlement to the company. In line with these results, in Experiment 3, the combined self-orientation of negotiators correlated .52 with favorability of settlement to the company for negotiators who planned strategies alone and .22 for negotiators who planned strategies in groups. Yet, the correlation was −.63 for negotiators who studied alone and .07 for those who studied in groups beforehand.

Differences in the self-orientation of negotiations were not related in any consistent way with negotiating outcomes. Likewise, neither the combined interaction-orientation scores for each pair of negotiators nor the differences between them, showed consistent correlations with any negotiation outcomes.

Postsession Critique

Pairs who had deadlocked and those who reached agreement quickly were queried about the reasons for their outcomes.

Some deadlocked partisans said they were unconcerned about the length or cost of the strike. Others argued that there was no reason to settle for less than the going rate since they assumed that workers could get jobs elsewhere.

One told of his negotiating process which seemed prone for failure. He would only negotiate on a one-for-one basis and would not give anymore than he felt he received in swapping issue by issue.

Two negotiators, who deadlocked and sat back-to-back in stony silence for the last part of the negotiating session, did not need to comment about their emotional involvement; but another deadlocked partisan felt he had been carried away by the role and had behaved completely realistically.

Early success in attaining what is regarded as an important concession and concern about the cost of the strike produced early settlement.

> . . . we agreed almost immediately on a 6-cents-an-hour increase, which my group had figured was the most important matter. On the less important items, I put a premium on time and did not worry about pennies. I figured it was better to remain below the going rate but settle the strike quickly.

An agreeable opponent also helped:

> . . . we started out by going through the few things that I was going to really build up as something big to trade on, and he went right through (the list of issues) and just gave them to me right away and so I just let him keep going. . . .

One pair seems to have exemplified Osgood's (1962) gradualism without knowing it. First, one negotiator made a small concession. This was followed by one from the other and so on down the line until full agreement was quickly reached. Yet, an astute negotiator commented:

> One thing I noticed (with strategy groups) was you could begin to detect after a while what their strategy was. For example, they would concede the smaller issue and skip over the most important, wages, and then hopefully come back later and use the argument, "oh, since we gave you that, how about. . . ."

Postsession Interviews

It was possible to complete individual postexperiment interviews with 39 of the 46 negotiators who had planned strategies for Experiment 3.

The two most popular methods of approaching the problem of planning strategies were either first to rate the issues for trading off ($N = 10$) or first to compare issues with community norms ($N = 9$). Four solitary planners began by rating issues according to costs, but no groups did this.

Of the 32 who reported having a primary goal in mind, 11 of the 15 union planners were aiming to maximize monetary gains, while only 7 of the 17 company planners were primarily out to minimize monetary costs. Eleven of 12 solitary planners with primary goals were concentrating on monetary gain while only 7 of 20 group planners focused first on money. Other primary goals distributed relatively evenly among the various planners included a settlement near the community average, bettering long-range relations between parties and immediately ending the strike.

While 15 of 16 company negotiators indicated that they had consciously considered what the union's goals might be, only 6 of 13 union negotiators had considered the company's aims. Twelve of 17 planning alone worried about what the other side was after; again, 9 of 12 in groups reported considering what the other side's goals might be.

Six company negotiators saw themselves pushing for a package deal; only one union man took this bargaining approach. Seven union negotiators said they had worked to trade off issues; only four company men were interested in applying this strategy. Other less frequently employed strategies included: arranging a priority of issues to be discussed ($N = 6$): minimax solutions ($N = 4$); pursuing one issue at a time ($N = 2$); finding the other party's opinion first ($N = 2$) and attacking the other side ($N = 1$).

Only 5 union and 1 company negotiator of the 39 were definitely convinced that the nonmonetary issues were very important.

While two-thirds of company negotiators felt committed to their prebargaining strategy, only half the union respondents were so committed. As might be expected, two of every three strategists coming from groups felt such commitment while only one of every two strategists who had planned alone felt similarly committed. Approximately the same ratios held in response to whether the actual bargaining proceeded according to plan or had to be revised. While 13 of 15 company negotiators would use the same strategy again, only 7 of 12 union bargainers would do so. Among those who had planned alone, there was a 12 to 1 preference to shift to planning strategies in groups while of those who had planned in groups, only half would prefer to switch to planning strategies alone.

CONCLUSIONS

Strategy versus Study Experience

The variety of contract outcomes points to the importance of focusing on how negotiators prepare themselves in advance of bargaining. Negotiators who had planned strategies in groups rather than studied the issues in groups were more likely to deadlock in the face of deadlines, but they could also achieve speedy settlements if the strategies they formulated happened to overlap with those of their counterparts.

Speedy settlements were not necessarily good settlements. In fact, if Campbell's (1960) criterion is accepted that the quality of the settlement is given by the closeness of the outcome to community norms, then speed was inversely related to quality under all 12 treatments. In all treatments, negotiating time was negatively correlated with departure of settlement from going rates in the community.

More tightly controlled experiments can be effected here, possibly yielding a better and simpler criterion of quality by eliminating the nonmonetary issues from the negotiations and suggesting nonoverlapping monetary goals to the conflicting parties, that is, \$60,000 gain for union, \$35,000 cost to company. (A pilot study with 85 contracts negotiating only monetary issues yielded a mean settlement of \$49,638.)

In comparison with settlements by those who had studied the issues, settlements by strategists departed more widely from the going rates in the first two but not the third experiment. Prenegotiation study favored better settlements for the company in the first and third, but not the second experiment. However, in all three experiments, strategists tended to agree more with their opponents in the rank importance of the issues than did those who studied the issues unilaterally in advance. Again, the complex outcomes appear to call for simplifying required negotiations. Nevertheless, maximum agreement on the relative importance of the issues can be affected by providing prenegotiation experience in bilateral study groups with one's future opponents and avoiding bilateral study groups with counterparts with whom one will not subsequently have to negotiate. And such agreement in advance seemed to speed negotiations somewhat, although it had no consistent association with the direction of the settlement.

There were relatively few very specific differences resulting from prenegotiation

experience in study or strategy planning in groups. Studying in groups seemed to result in more decisions for departmental rather than plant-wide seniority and greater importance being attached to the hospital and medical plan. Strategy planning resulted in greater importance being assigned to the proposed union representative on the company board.

Rather than pursue this question about the differential effects of study rather than strategy on contract outcomes, our mixed bag of results and the critiques that follow suggest that it may be more profitable to move into an examination of specific tactics which may increase the speed of contract resolution and/or enhance the quality of outcomes. Requiring resolution of only the five monetary issues, different samples of prenegotiation strategists can be asked to plan in groups as follows:

1. Aim to obtain a settlement at the community norm
2. Minimize cost (or maximize gain)
3. Develop tactics on estimates of the opposing party's goals
4. Use as a strategy:
   a. Package deal
   b. Trade offs
5. Begin negotiating with:
   a. The most important issue
   b. The least important issue
   c. By a search of the opposition's views.

Group versus Individual Prenegotiation Preparation

Group commitment cumulated with strategic thinking to hold bargainers to the longest decision times. Group experience produced deadlocks which never occurred as a consequence of individual preparation. Opposing parties who had met beforehand in unilateral groups were in greater agreement about the importance of the issues (yet took longer to reach contractual agreement) than those who had prepared alone. Those who prepared in groups felt they had more defensible positions than those who prepared individually. Most other specific variables were not appreciably affected although group preparation raised the saliency of the hospital-medical plan and lowered the perceived importance of the seniority question.

In short, the reinforcing properties of group experience were revealed. As a general procedure, opposing positions of negotiators can be readily hardened by prenegotiation reinforcement in unilateral groups.

Personal Orientation

When task-orientation is high among negotiators, settlements are closer to going rates—which may be interpreted as illustrative of high-quality resolution; but when one negotiator is high and the other low in task orientation, the reverse is true about contract outcomes. It would seem profitable to follow up this finding with a replication with a larger sample and with a more simplified set of negotiations to yield a simpler monetary criterion of quality for contract outcomes.

Such a simpler negotiation problem has been formulated. Results obtained with it will be the subject of subsequent reports.

APPENDIX

TABLE A1

ANALYSES OF VARIANCE OF THE EFFECTS ON CONTRACT OUTCOMES OF WHETHER UNION AND/OR COMPANY PLANNED STRATEGIES OR STUDIED ISSUES BEFORE NEGOTIATING (EXPERIMENT 2)

Dependent variable	Treatment			Error	Total variance
	(i) Strategy versus study effect due to union	(j) Strategy versus study effect due to company	(i × j) Interaction effect		
	1 *df*	1 *df*	1 *df*	44 *df*	47 *df*
Duration of strike[a]					85.44
Mean squares	32.01	24.65	10.08	9.51	
F ratio	3.36	2.59	1.06		
Departure from going rate					547881.
Mean squares	6984.	581.	8086.	12096.	
F ratio	.577	.005	.668		
Settlement favors company					400182.
Mean squares	2146.	62424.	196.	7623.	
F ratio	.282	8.189**	.025		
Agreement on issues					5.098
Mean squares	.3798	.0150	.0728	.1052	
F ratio	3.61	.143	.692		

[a] These data exclude the deadlocks. When deadlocks are included in the analyses, converted to reciprocal of time with the assumption that the deadlock was infinitely long in duration so that $1/\infty = 0$, the attained mean variance in harmonic mean time among the four treatments was significant at the 5% level.

** $p < .01$.

TABLE A2

ANALYSES OF VARIANCE OF THE EFFECTS ON SPECIFIC SETTLEMENTS OF WHETHER UNION AND/OR COMPANY PLANNED STRATEGIES OR STUDIED ISSUES BEFORE NEGOTIATING (EXPERIMENT 2)

Dependent variable	Treatment: (i) Strategy versus study effect due to union	Treatment: (j) Strategy versus study effect due to company	Treatment: (i × j) Interaction effect	Error	Total variance
	1 *df*	1 *df*	1 *df*	44 *df*	47 *df*
Hospital-medical					17.67
Mean squares	.08	.00	2.08	.35	
F ratio	.23	.00	5.91*		
Wages					45.98
Mean squares	.19	.52	3.52	.94	
F ratio	.20	.55	3.71		
Sliding pay					11.25
Mean squares	.75	.33	.33	.22	
F ratio	3.36	1.49	1.49		
Seniority					11.25
Mean squares	.08	.75	.75	.22	
F ratio	.38	3.4	3.4		
Union representative[a]					
Night-shift differential					155.81
Mean squares	22.69	11.02	1.69	2.73	
F ratio	8.29**	4.03*	.62		
Vacation pay					29.98
Mean squares	.02	1.69	1.02	27.25	
F ratio	.03	2.72	1.65		
Work rules committee					11.67
Mean squares	.08	2.08	.33	.21	
F ratio	.40	10.00**	1.60		
Checkoff					11.81
Mean squares	.52	.02	.18	.25	
F ratio	2.07	.08	.74		

[a] No variance in response as a function of treatment.
* $p < .05$.
** $p < .01$.

TABLE A3

ANALYSES OF VARIANCE OF THE EFFECTS ON SATISFACTION WITH CONTRACTS AS A FUNCTION OF WHETHER UNION AND/OR COMPANY PLANNED STRATEGIES OR STUDIED ISSUES BEFORE NEGOTIATING (EXPERIMENT 2)

Dependent variable	Treatment: (i) Strategy versus study effect due to union	Treatment: (j) Strategy versus study effect due to company	Treatment: (i × j) Interaction effect	Error	Total variance
	1 *df*	1 *df*	1 *df*	92 *df*	95 *df*
Preferred length?					16583.
Mean squares	5133.	63.37	.3750	123.7	
F ratio	41.47**	.512	.003		
Other negotiator's performance?					33.98
Mean squares	.0938	1.260	2.344	.3293	
F ratio	.285	3.828	7.118**		
Acceptable contract?					30.62
Mean squares	.3750	2.0417	.0417	.3062	
F ratio	1.225	6.668*	.1360		
Who fared better?					27.24
Mean squares	.5104	.0938	.2604	.2867	
F ratio	1.780	.3270	.9080		
Defensible position?					37.83
Mean squares	6.000	.1667	.0000	.3442	
F ratio	17.43	.4840	.0000		
Congruence of role?					71.74
Mean squares	19.26	1.260	.2604	.5539	
F ratio	34.77**	2.276	.470		

* $p < .05$.
** $p < .01$.

TABLE A4

ANALYSES OF VARIANCE OF THE EFFECTS ON JUDGED RANK IMPORTANCE OF ISSUES AS A FUNCTION OF WHETHER UNION AND/OR COMPANY PLANNED STRATEGIES OR STUDIED ISSUES BEFORE NEGOTIATING (EXPERIMENT 2)

Dependent variable	Treatment			Error	Total variance
	(i) Strategy versus study effect due to union	(j) Strategy versus study effect due to company	(i × j) Interaction effect		
	1 *df*	1 *df*	1 *df*	92*df*	96 *df*
Hospital-medical					243.9
Mean squares	5.042	.3750	1.042	2.581	
F ratio	1.953	.145	.403		
Wages					125.4
Mean squares	6.510	.5104	.2604	1.284	
F ratio	5.07*	.398	.203		
Sliding pay					419.9
Mean squares	44.01	17.51	1.260	3.881	
F ratio	11.34**	4.51*	.325		
Seniority					429.6
Mean squares	51.04	12.04	.3750	3.980	
F ratio	12.82**	3.02	.094		
Union representative					867.2
Mean squares	420.8	3.01	15.84	4.647	
F ratio	90.6**	.648	3.41		
Night-shift differential					270.7
Mean squares	82.51	.0937	1.260	2.031	
F ratio	40.6**	.046	.620		
Vacation pay					287.2
Mean squares	14.26	.5104	17.51	2.771	
F ratio	5.14*	.184	6.32*		
Work rules committee					370.2
Mean squares	.5104	36.26	1.260	3.611	
F ratio	.141	10.04**	.3490		
Checkoff					197.6
Mean squares	13.50	.6667	.3750	1.990	
F ratio	6.78*	.335	.188		

* $p < .05$.
** $p < .01$.

TABLE A5

ANALYSIS OF VARIANCE OF THE EFFECTS ON CONTRACT OUTCOMES OF WHETHER UNION AND COMPANY PLANNED STRATEGIES OR STUDIED ISSUES IN GROUPS OR ALONE BEFORE NEGOTIATING (EXPERIMENT 3)

Dependent variable	Treatment			Error	Total variance
	(i) Study versus strategy	(j) Group versus alone	(i × j) Interaction		
	1 *df*	1 *df*	1 *df*	32 *df*	35 *df*
Duration of strike[a]					351.79
Mean squares	13.44	31.21	44.10	10.75	
F ratio	1.25	2.90	2.52		
Departure from going rate					307271.
Mean squares	35094.	841.	14240.	8034.	
F ratio	4.37*	.105	1.77		
Settlement favors company					483510.
Mean squares	5244.	2093.	203.	1275.	
F ratio	4.11	1.64	.159		
Agreement on issues					3.649
Mean squares	.0072	.1034	.0078	.1103	
F ratio	.065	.937	.071		

[a] See Footnote a of Table A1. Here the treatment effect was significant at the 1% level when harmonic means were calculated.

* $p < .05$.

TABLE A6

ANALYSIS OF VARIANCE OF THE EFFECTS ON SPECIFIC SETTLEMENTS OF WHETHER UNION AND COMPANY PLANNED STRATEGIES OR STUDIED ISSUES IN GROUPS OR ALONE BEFORE NEGOTIATING (EXPERIMENT 3)

Dependent variable	Treatment			Error	Total variance
	(i) Study versus strategy	(j) Group versus alone	(i × j) Interaction		
	1 *df*	1 *df*	1 *df*	32 *df*	35 *df*
Hospital-medical					10.75
Mean squares	.69	.03	.03	.31	
F ratio	2.23	.97	.97		
Wages					81.56
Mean squares	.11	4.00	1.00	2.39	
F ratio	.05	1.67	.42		
Sliding pay					8.56
Mean squares	.11	.11	.11	.26	
F ratio	.42	.42	.42		
Seniority					8.56
Mean squares	1.00	.11	1.00	.20	
F ratio	5.00*	.55	5.00*		
Union representative[a]					
Night-shift differential					93.64
Mean squares	6.94	6.94	8.03	2.63	
F ratio	2.64	2.64	3.05		
Vacation pay					26.75
Mean squares	.25	.03	.69	.81	
F ratio	.31	.04	.85		
Work rules committee					7.64
Mean squares	2.25	.03	.03	.17	
F ratio	13.24**				
Checkoff					8.56
Mean squares	.00	.44	1.00	.22	
F ratio	.00	2.00	4.55*		

[a] No variance as a fraction of treatment.
* $p < .05$.
** $p < .01$.

TABLE A7

ANALYSIS OF VARIANCE OF THE EFFECTS ON SATISFACTION WITH CONTRACTS AS A FUNCTION OF WHETHER UNION AND COMPANY PLANNED STRATEGIES OR STUDIED ISSUES IN GROUPS OR ALONE BEFORE NEGOTIATING (EXPERIMENT 3)

Dependent variable	Treatment							Error	Total variance
	(i) Study versus strategy	(j) Group versus alone	(k) Union versus company respondent	Interactions					
				(i × j)	(i × k)	(j × k)	(i × j × k)		
	1 *df*	1 *df*	1 *df*	1 *df*	1 *df*	1 *df*	1 *df*	72 *df*	79 *df*
Preferred length?									13736
Mean squares	2599	460.8	583.2	405.0	520.2	16.2	583.2	119.0	
F ratio	21.8**	3.87	4.90*	3.40	4.37*	.14	4.90*		
Other negotiatior's performance?									25.55
Mean squares	.050	.450	1.250	.450	1.250	1.250	1.250	.272	
F ratio	.18	1.65	4.60*	1.65	4.60*	4.60*	4.60*		
Acceptable contract?									43.89
Mean squares	.0125	.1125	2.1125	.6125	.0125	.0125	1.5125	.5486	
F ratio	.023	.205	3.85	1.12	.023	.023	2.76		
Who fared better?									30.89
Mean squares	.0125	.0125	1.5125	.1125	2.1125	.3125	.1125	.3708	
F ratio	.034	.034	4.08	.30	5.72*	.843	.30		
Defensible position?									42.75
Mean squares	2.4500	1.2500	2.4500	.0500	.0500	1.2500	.0500	.4889	
F ratio	5.01*	2.56	5.01*	.102	.102	2.56	.102		
Consequence of role?									46.80
Mean squares	3.2000	.8000	.0000	.2000	.0000	.2000	.8000	.5778	
F ratio	5.54*	1.384	.00	.346	.00	.346	1.384	.346	

* $p < .05$.
** $p < .01$.

TABLE A8

ANALYSIS OF VARIANCE OF THE EFFECTS ON JUDGED RANK IMPORTANCE OF ISSUES AS A FUNCTION OF WHETHER UNION AND COMPANY PLANNED STRATEGIES OR STUDIED ISSUES IN GROUPS OR ALONE BEFORE NEGOTIATING (EXPERIMENT 2)

Dependent variable	Treatment							Error	Total Variance
	(i) Study versus strategy	(j) Group versus alone	(k) Union versus company respondent	Interactions					
				(i × j)	(i × k)	(j × k)	(i × j × k)		
	1 df	1 df	1 df	1 df	1 df	1 df	1 df	72 df	79 df
Hospital-medical									288.0
Mean squares	7.220	.1800	1.280	.0450	.6050	2.645	.0800	.2325	
F ratio	31.05**	.77	5.51*	.19	2.60	11.38**	.34		
Wages									62.0
Mean squares	.3125	.6125	.6125	.3125	3.613	6.613	.0125	.6931	
F ratio	.45	.88	.88	.45	15.5	9.54**	.02		
Sliding pay									284.9
Mean squares	55.13	10.13	36.13	3.125	1.125	36.13	36.13	32.10	
F ratio	1.72	.32	1.12	.10	.35	1.12	1.12		
Seniority									300.2
Mean squares	11.25	11.25	60.50	7.200	.2000	24.20	.050	3.333	
F ratio	3.38	3.38	1.82	2.16	.06	7.26**	.02		
Union representative									539.4
Mean squares	137.8	.6125	2.813	2.813	7.813	1.513	5.513	5.285	
F ratio	26.07**	.01	.53	.53	1.48	.29	1.04		
Night-shift differential									223.8
Mean squares	39.20	.8000	1.800	9.800	.000	7.200	.2000	2.289	
F ratio	17.13**	.35	.79	4.28	0.0	3.15	.09		
Vacation pay									169.5
Mean squares	2.813	.0125	.6125	.1125	.1125	3.613	9.112	2.126	
F ratios	1.32	.01	.29	.05	.05	1.70	4.29		
Work rules committee									435.0
Mean squares	52.81	5.513	3.613	4.513	15.31	12.01	23.11	4.432	
F ratios	11.92**	1.24	.82	1.02	3.45	2.71	5.21		
Checkoff									211.6
Mean squares	5.000	3.200	4.050	.0500	.8000	.2000	8.450	2.636	
F ratios	1.90	1.21	1.54	.02	3.04	.76	3.21		

* $p < .05$.
** $p < .01$.

REFERENCES

AMERICAN MANAGEMENT ASSOCIATION. Workshop Seminar No. 8172-12. New York, June 17–19, 1963.

BASS, B. M. *The orientation inventory.* Palo Alto: Consulting Psychologists Press, 1962.

BLAKE, R. R., & MOUTON, J. S. Competition, communication and conformity. In I. A. Berg & B. M. Bass (Eds.) *Conformity and deviation.* New York: Harper, 1961.

CAMPBELL, R. J. Originality in group productivty. III. Partisan commitment and productive independence in a collective bargaining situation. The Ohio State University Research Foundation, Columbus, Ohio, 1960.

BOULDING, K. E. *Conflict and defense: A general theory.* New York: Harper, 1962.

DEUTSCH, M. Conditions affecting cooperation. I. Factors related to the initiation of cooperation.

II. Trust and cooperation. *Final Technical Report* Nonr 285(10). Research Center for Human Relations, New York University, 1957.

Gouldner, A. W. *Wildcat strike.* Yellow Springs, Ohio: Antioch Press, 1954.

Hepler, J. W. The relationship between the efficiency of the group decision-making process and group polarization. Unpublished doctoral dissertation, Ohio State University, Columbus, 1953.

Loomis, J. L. Communication and the development of trust. Contract Nonr 285(10). Research Center for Human Relations, New York University, 1957.

Mack, R. W., & Synder, R. C. The analysis of social conflict—toward an overview and synthesis. *Journal of Conflict Resolution,* 1957, 1, 212–248.

Osgood, C. E. Psychological concepts in arms control and/or graduated unilateral disarmament. In L. Carter (Chm.) Arms control and the psychologist. Symposium presented at American Psychological Association, St. Louis, September 1962.

Rapaport, A. *Fights, games and debates.* Ann Arbor: University of Michigan Press, 1960.

Schelling, T. M. Bargaining, communication and limited war. *Journal of Conflict Resolution,* 1957, 1, 19–36.

Scodal, A., Minas, J. S., Ratoosh, P., & Lipetz, M. Some descriptive aspects of two person non-zero sum games. *Journal of Conflict Resolution,* 1959, 3, 114–119.

Sherif, M., et al. *Intergroup conflict and cooperation. The Robber's Cave Experiment.* Norman: University of Oklahoma Book Exchange, 1961.

Siegel, S., & Fouraker, L. E. *Bargaining and group decision-making.* New York: McGraw-Hill, 1960.

Vegas, O. V., Frye, R. L., & Cassens, F. P. Learning set as a determinant of perceived cooperation and competition. In D. C. Gilbert (Chm.) Symposium presented at American Psychological Association, Los Angeles, 1964.

(Received October 19, 1965)

Vol. 80, No. 7 Whole No. 615, 1966

Psychological Monographs: General and Applied

THE CLASSIFICATION OF CHILDREN'S PSYCHIATRIC SYMPTOMS:

A FACTOR-ANALYTIC STUDY[1]

THOMAS M. ACHENBACH[2]

University of Minnesota

Symptoms from the case histories of 300 male and 300 female child psychiatric patients were analyzed, separately for each sex, by the principal-factor method with quartimax, varimax, and oblimin rotations. Classification of the Ss according to the 1st principal factor and the reliable rotated factors showed that symptom clusterings at 2 levels of generality were present: there was a general polar dichotomy given the label Internalizing versus Externalizing, and there were several specific syndromes, some resembling traditional psychiatric diagnoses and some peculiar to certain developmental stages. Biographical differences among the Ss suggested that the Internalizing-Externalizing dichotomy and those specific syndromes subsumed by it reflected differences in socialization, while the syndromes not subsumed by it did not reflect socialization differences. The factors obtained can be used to classify child psychiatric patients for research purposes.

The advent of modern psychiatry was heralded in part by the introduction in 1883 of Kraepelin's diagnostic system. In its earliest forms, this system rested on the assumption that all mental disorders were due to physical pathology in the brain. Kraepelin's goal was to devise categories of disorder based upon descriptions of the symptom syndromes manifested and upon the observed courses of illness. It was expected that medical research would eventually uncover different physical etiologies for the different categories thus formulated.

Although Kraepelin himself soon focused his interest upon psychological processes and appears by 1900 to have abandoned the dogma that all mental disorders were due to brain dysfunction, the medical disease model has continued to represent one possible format for psychiatric diagnosis. It would appear that, for mental disorders having some specific physical agent or malfunction as the necessary, sine qua non, cause, the orthodox disease model is indeed the appropriate one. Like other organic diseases, such disorders should ultimately be defined and treated with reference to the physical etiology.

A second possible model for some categories of mental disorder is illustrated by the conjectures of Rado (1953) and Meehl (1962) on the nature of schizophrenia. According to this model, certain inherited or constitutional anomalies are the necessary but not sufficient causes of the clinical disorder. Given the prerequisite anomaly, the typical social learning regime will result in a personality organization which is suscep-

[1] This research was conducted while the author was a National Institute of Mental Health Predoctoral Fellow. It was supported by National Institute of Mental Health Research Grant MH-06170 (Norman Garmezy, Principal Investigator), and National Institute of Mental Health Research Grant MH-06809 (Edward Zigler, Principal Investigator).

[2] The author would like to express his thanks to the following people: Norman Garmezy for his support and guidance throughout the course of the study; Reynold Jensen and Wentworth Quast of the University of Minnesota Hospital Child Psychiatry Unit for helpful advice and for making the case histories available; Edward Zigler for providing the author with the opportunity to work on the study during the summer of 1964 and for his critical reading of the manuscript; Beverly Kaemmer for her help in reading case histories and at other stages in the research; and Conrad Katzenmeyer for his extensive help in the factor analyses. All computer work was carried out at the University of Minnesota Numerical Analysis Center.

The author is now at Yale University.

tible to clinical schizophrenia. Given other constitutional weaknesses and/or serious psychological stresses, the schizotypic personality can potentially manifest a clinical schizophrenic reaction. The disorder is thus the product of the interaction between a specific constitutional attribute, the learning history resulting when an individual with such a constitutional attribute is subjected to the usual environmental regime, and the presence of precipitating stress.

A third possible model for mental disorders is that of the behavior theorists. They (e.g., Bandura, 1964; Bandura & Walters, 1963) maintain that behavioral deviations should be regarded as learned reactions rather than as symptomatic "disease" manifestations. According to this position, symptomatic behavior is to be explained in terms of the social conditioning of the specific behavior observed, rather than by recourse to medical analogies.

It is conceivable that these three models validly represent three different types of mental disorder. In order to discover categories of disorders for which each of the types of model might be appropriate, it would be useful to have a classification system which provides operationally defined categories and which can at the same time play a heuristic role for further research. This would be especially true for disorders having no known physical etiology, since those for which a physical etiology can now be identified are adequately defined by that etiology in terms of the disease model.

Disorders without a known physical etiology are currently lumped together in the "functional" category. Within this category, psychiatric usage tends to combine the disease "sign" approach of organic diagnosis with behavioral and dynamic concepts in confusing and inconsistent ways. Classificatory research is needed to differentiate functional disorders according to different conceptual models. The isolation of clusters of empirical attributes is a logical first step in this direction, and this has frequently been approached by intercorrelating psychiatric symptoms. Among the most extensive studies of this kind[8] was that by Wittenborn, Holzberg, and Simon (1953). With the help of several psychiatrists, they constructed 55 symptom scales, each of which contained three or four objectively described behaviors. The symptoms occurring in a sample of state hospital patients were recorded by psychiatrists and factor analyzed by the centroid method, with an orthogonal rotation. Nine clusters, all similar to traditional diagnostic categories, were found and were given the following labels: acute anxiety, conversion hysteria, manic state, depressed state, schizophrenic excitement, paranoid condition, paranoid schizophrenic, hebephrenic schizophrenic, and phobic compulsive.

With a sample of psychiatric referrals which included some jail and outpatient cases as well as state mental hospital patients, Phillips and Rabinovitch (1958) used a different technique for isolating symptom clusters and obtained results different from those of Wittenborn. They checked the presence or absence of 46 different presenting symptoms listed by a psychiatrist or referring physician in the case histories of patients. By grouping together all symptoms which were shown by chi-square analysis to be positively interrelated, they found three symptom clusters. Symptoms not statistically related to the clusters were assigned by rational analysis and the clusters were interpreted as representing patterns of (*a*) avoidance of others; (*b*) self-indulgence and turning against others; and, (*c*) self-deprivation and turning against the self. These three clusters were confirmed in a new sample of case histories.

Guertin (1952) recorded the presence or absence of 77 symptoms in each of 100 hospitalized schizophrenics. An unrotated centroid analysis of the 52 most frequent symptoms produced six factors. These were labeled Excitement-Hostility, Retardation and Withdrawal, Guilt-Conflict, Confused Withdrawal, Persecutory-Suspicious, and Personality Disorganization. Except for the presence of psychotic symptoms in all of

[8] Symptom studies which were designed to test specific hypotheses (e.g., those of Lorr and his colleagues, cf. Lorr, Klett, & McNair, 1963), or which attempted other than empirical analyses of a broad range of functional symptomatology will not be considered here.

them and the greater specificity of symptom categories, the first three factors resemble the three clusters found by Phillips and Rabinovitch.

Empirical studies of child symptom groupings have followed somewhat similar lines. Hewitt and Jenkins (1946) recorded the presence or absence of 94 symptomatic traits occurring in each of 500 child guidance clinic case histories. Forty-five of the 94 traits, chosen either because of high frequency or "obvious clinical importance," were cross-tabulated with the whole series of traits. By inspection, three clusters of traits resembling three behavior syndromes previously suggested by a committee of consultants were found. From the 10 to 12 traits in each cluster, 6 or 7 were chosen to form the final three clusters by which cases were to be classified. For each cluster, traits were chosen which correlated at least .30 with most of the other traits in the cluster and which fit the clinical picture suggested by the cluster. The clusters thus formed were interpreted as representing the Overinhibited Child, the Unsocialized Aggressive Child, and the Socialized Delinquent Child.

In a recent study, Dreger (Dreger, 1964; Dreger, Lewis, Rich, Miller, Reid, Overlade, Taffel, & Flemming, 1964) had parents or parent surrogates of child clinic patients sort 229 discrete behavioral items according to whether or not the child had manifested them in the previous 6 months. The intercorrelations of 142 behavior items and demographic variables were analyzed by the principal-factor method with an oblimin rotation of the 10 largest factors. The rotated factors were left unlabeled, but their descriptions suggest that some of them resemble the general clusters found in the Phillips-Rabinovitch, Guertin, and Hewitt-Jenkins studies, while others vaguely resemble discrete syndromes like those found in the Wittenborn study. A cluster analysis of factor scores revealed five types of clinical cases, one of which (Egocentric Antisocial Aggressiveness) appears to match Hewitt and Jenkins' Unsocialized Aggressive Child, and two others of which (Relatively Immature, Nonsociable, Semisurgent Egocentricity and Sociable Anxiety) appear to resemble Hewitt and Jenkins' Overinhibited Child.

Despite the theoretical emphasis, since Freud, upon the continuity between problems of childhood adjustment and the occurrence of adult disorders, there has been remarkably little synthesis in the study of the forms of child and adult disorders. The American Psychiatric Association's *Diagnostic and Statistical Manual of Mental Disorders* (1952) is relatively undifferentiated in the area of childhood disorders, listing only the following categories: Adjustment Reaction of Infancy; Adjustment Reaction of Childhood, with the subcategories of Habit Disturbance, Conduct Disturbance, and Neurotic Traits; Adjustment Reaction of Adolescence; and, Schizophrenic Reaction, Childhood Type. While the Adjustment Reaction categories reflect the theoretical conception that such disturbances are transient and peculiar to certain developmental periods, the practical result is that many child clinics, for lack of more specific categories, find it necessary to assign large proportions of their functional cases to these categories, with little gain in utility through such assignments. In child psychiatric research and theory, distinctions like that between Adjustment Reaction with Conduct Disturbance and Adjustment Reaction with Neurotic Traits appear to be among the most frequently employed. Ross' (1959) textbook, for example, makes a fundamental distinction between the child manifesting aggressive behavior and the child manifesting withdrawn behavior and physical symptoms. A more theoretical version of this distinction is implied in Bard, Sidwell, and Wittenbrook's (1955) proposal that children be classified Healthy, Asocial, Psychoneurotic, or Antisocial, according to the kind and degree of standards they had introjected.

Although the empirical isolation of clusters of attributes may be a necessary first step in some sciences, a more advanced step is to achieve a classification system based upon theoretical principles which are assumed to determine the observed groupings. The proposal of Bard, Sidwell, and Wittenbrook that children be classified according to the degree and kind of standards they

had introjected illustrates classification dictated by a theoretical principle. Such a classification could be regarded as an advance over purely empirical classification if it provides reliably discriminable categories which convey a theoretical understanding of the phenomena and which can eventually guide differential action on the part of those using the categories. The three symptom clusters found by Hewitt and Jenkins for children might be regarded as operational analogues of the distinction proposed by Bard, Sidwell, and Wittenbrook. Similarly, the "self-indulgence and turning against others," and "self-deprivation and turning against the self" symptom clusters found by Phillips and Rabinovitch appear to reflect the same distinction between adult patients who have and adult patients who have not introjected social standards. The Excitement-Hostility and Guilt-Conflict factors found by Guertin would seem to indicate a similar distinction among schizophrenic patients. It might thus be concluded that there is at least some similarity in the general symptom patterns of childhood and adult functional disorders, and that the distinction between these patterns involves, according to one interpretation, the presence or absence of introjected social standards. The value of this distinction has been demonstrated for child disorders by Hewitt and Jenkins and by Bennett (1960). Both studies found that the "overinhibited" or "neurotic" children tended to come from stable homes while the "aggressive" and "delinquent" children tended to come from broken homes where the parents presented many social problems. The value of the same distinction was demonstrated for adult disorders by Zigler and Phillips (1960). They found that symptoms from the "self-deprivation and turning against the self" cluster were positively related to high standing on the social effectiveness variables of age, marital status, employment history, occupation, education, and intelligence. Symptoms from the "self-indulgence and turning against others" cluster were inversely related to standing on the social effectiveness variables. Furthermore, women were found to manifest more symptoms of "self-deprivation and turning against the self," whereas men were found to manifest relatively more symptoms of "self-indulgence and turning against others." Zigler and Phillips (1960) suggested "that both social effectiveness and patterns of symptomatic behavior represent the degree to which an individual tends either to resist or to conform to the mores of society. [p. 237]."

Beside acknowledging the parallel findings of the major distinction between clusters of "intropunitive" and "extrapunitive" symptoms for children and adults in the Hewitt-Jenkins, Phillips-Rabinovitch, and Guertin studies, it is difficult to reconcile specific inconsistencies among the empirical studies, to evaluate the degree of child-adult continuity beyond the one evident distinction, and to evaluate the relation of the traditional psychiatric syndromes, like those found by Wittenborn, to the more general symptom clusters. Although not necessarily invalidating their findings, certain features of the studies prevent conclusive evaluations of the empirical purity of the clusters found. Any delineation of clusters of attributes must be influenced by the goals of the user and the methods employed. In attempting either a theoretical synthesis or further empirical work, it is well to consider sources of inconsistency which cannot be simply written off as error, but which must be attributed to the multifaceted nature of the classificatory enterprise.

First, it would appear that the formation of groupings found by Wittenborn may have been influenced by traditional psychiatric stereotypes. The psychiatrists constructing the rating scales as well as those making the symptom ratings were likely to have been trained to expect certain symptoms to occur in the presence of certain other symptoms and to be more alert to the occurrence of symptoms which coincided with the categories to which they were accustomed. Although Wittenborn included psychiatrists of different theoretical persuasions, there may well have been sufficient overlap among them at the level of descriptive diagnosis to make this a significant source of bias. If this were the case, the discovery of factors resembling traditional diagnostic categories would not be surprising. Moreover, Witten-

born sought factor rotations which were "reminiscent of classical diagnostic concepts. [Wittenborn, 1957, p. 445]."

Guertin obtained his symptom ratings largely through interviews with the patients. The type of probing and the symptoms thus sought and elicited might be expected similarly to reflect the biases of psychiatric practice. The Phillips-Rabinovitch findings may also have been subject to this source of bias because the reports of a physician or the examining psychiatrist provided the data. However, the type of statistical analysis they employed was less likely than a factor analysis to produce syndromes as discrete as the traditional ones. A second reservation about the empirical purity of the Phillips-Rabinovitch clusters is that some symptoms were assigned rationally, rather than on the basis of observed statistical relationships.

Dreger's study suffers weaknesses which make it difficult to compare to the other studies. While its findings were probably not biased by psychiatric stereotypes, the interrater agreement between parents or parent surrogates rating the same child ranged from only 10% to 55%, with a mean of 36%. Moreover, nonpsychiatric control subjects (*S*s) were found to exhibit proportionally more sadistic behaviors than were reported for the clinic children, suggesting a possible bias in the ratings by the clinic children's parents. Dreger's use of very specific "first order" behavior items rather than more general symptom categories may have increased objectivity somewhat, but at the same time probably resulted in a low frequency of responding in many categories and produced clusterings of items at a level too molecular to allow useful interpretation.

Finally, the Hewitt-Jenkins study included several highly similar symptom categories which may explain the emergence of the separate Socialized Delinquent and Unsocialized Aggressive clusters not found in other studies. For example, two of the six symptoms correlating most highly within the Unsocialized Aggressive cluster were "assaultive tendencies" and "initiatory fighting." Among the seven symptoms correlating most highly within the Socialized Delinquent cluster were "bad companions," "gang activities," "cooperative stealing," "furtive stealing," "truancy from home," and "out late nights." It would appear that in many cases the same behavior could be assigned to each of two similar categories. Furthermore, as pointed out by Jenkins (1964), symptoms of developmental retardation, brain damage, and schizophrenia were not included in the symptom check list, so these syndromes were unlikely to appear in the statistical analysis. By using reports from all sources in the records, Hewitt and Jenkins may have minimized psychiatric biases at the observer level, but the formation of trait clusters appeared explicitly to involve much "clinical" judgment, and the clusters found conformed without exception to the judges' expectations.

A general consideration regarding most of the studies is that they employed heterogeneous subject populations and failed to make comparisons among homogeneous segments of their heterogeneous samples. Some of the studies included organic, retarded, racially mixed, and physically handicapped patients, and all of them included irregular age and sex distributions. It is sometimes possible that such correlational analyses of heterogeneous samples will obscure opposite tendencies which may be present for different subgroups within the samples. Aside from the general inconsistencies among the studies, comparison of the different results is made difficult by the different methods of analysis employed. The cluster analyses employed in the Phillips-Rabinovitch and Hewitt-Jenkins studies were likely to produce fewer and more general groupings than were the rotated factor methods employed by Wittenborn and Dreger. Guertin's use of a diagnostically homogeneous sample produced clusterings which would probably have been obscured in a similar factor analysis of a diagnostically heterogeneous sample.

Purposes of the Present Study

The primary purpose of the present study was to attempt to elucidate, in the child symptom domain, the relationship between the general symptom clusters found in the Hewitt-Jenkins, Phillips-Rabinovitch, and Guertin studies, and the specific functional

syndromes employed in adult psychiatry and found in the Wittenborn study. One obvious possibility is that the more specific syndromes are subtypes of the more general clusters. Lorr (1957) found evidence for this possibility when he subjected some of the Wittenborn data (Wittenborn & Holzberg, 1951) to factor analysis by oblique and second-order methods. He found three second-order factors, the first of which represented a "general turning against the self." The second represented a "paranoid belligerence . . . defined by paranoid ideation, combativeness, and motor restlessness." The third factor represented "thinking disorganization joined with slowed psychomotor activity [Lorr, Klett, & McNair, 1963, p. 83]." These are clearly reminiscent of the three clusters found by Phillips and Rabinovitch, and two of them are suggestive of the Overinhibited versus two extrapunitive clusters found by Hewitt and Jenkins. A maximally flexible factor-analytic methodology was to be employed in the present study to test this same possibility for the child-symptom domain. To establish conclusions with as much confidence as possible, the elimination or minimization of the limitations evident in earlier studies was also sought.

A second purpose of the study was to obtain, for research purposes, a more differentiated empirical classification of child psychiatric cases than is now available. As pointed out by Zigler and Phillips (1961), the value of such a purely descriptive classificatory schema would be much enhanced if useful class correlates beyond the defining characteristics of its classes could be discovered.

A third purpose was to classify individual cases according to the factors obtained in order to see to what extent they represented types of cases. If it were indeed found that both specific syndromes and more general clusterings appeared in the factor analyses, the classification of cases according to their resemblance to both types of factors could reveal the degree to which some factors represented subcategories of other, more general, factors.

A fourth purpose of the study was to examine readily available biographical data for relationships to the classifications derived from the factor analyses of symptoms. Statistically significant relationships between classification by the factors and standing on the biographical variables would indicate that the symptom categories validly discriminated cases at least in terms of those particular biographical items. Furthermore, comparisons with earlier studies which have investigated such biographical differences could be made.

The final purpose of the study was heuristic. The elucidation of the relationships between empirical groupings of symptoms at various levels of generality might suggest researchable hypotheses about the functional relationships determining those groupings. In conjunction with findings of significant differences on the biographical variables and comparisons with previous studies, intermediate level constructs, subject to further test, might be evolved to explain the symptom groupings obtained and to aid in the choice of diagnostic models for further research.

PROCEDURE

Data Collection

The search for empirical groupings of attributes presents several problems for the collection of data. Decisions as to the sources of observations, the level of abstraction at which observations are to be classified for analysis, and the role of the observers are necessary. Collecting the attributes of disordered behavior requires a human observer who must subjectively abstract from his experience and categorize his observations. In addition, the human observer is quite likely to influence the behavior manifested by the human *S* observed. The present study attempted to minimize the influence of systematic biases in the observer by using the reports of several different observers, having different roles with respect to *S*, as contained in case histories. It attempted to minimize systematic biases in the assignment of observations to categories for analysis by using raters who lacked specialized training in clinical practice and who would not be expected to share the clinical stereotypes as strongly as would trained practitioners. Finally, it attempted to reduce artifactual clusterings of attributes by employing a mutually exclusive system of ratings, whereby any reported observation was to be entered in only one category and where the categories were defined so as to minimize overlap as much as possible.

As for the level of abstraction of the rating categories, symptom categories were sought which were objective and required as little inference as possible, but which were not so specific as to pre-

clude meaningful abstraction due either to low frequencies or to overly molecular units. To attain this goal and to give some basis for comparison with earlier findings, a symptom check list was constructed from items which regularly appeared in previous studies, which seemed to involve minimal inference, which could be considered mutually exclusive with regard to specific observations, and which were not excessively molecular. In addition, 40 case histories at the University of Minnesota Hospital Child Psychiatry Unit were read to obtain further symptom categories. By this means and by adding a few new symptoms which occurred in the data samples, a list of 91 symptoms (Appendix A), to be checked if present, was constructed. In effect, the definition of "symptom" here approximated that put forth by Lorr et al. (1963, p. 3). It was not intended necessarily to refer to signs or tokens of internal disease, but only to deviant behaviors, postures, attitudes, or verbalizations generally accepted as reasons for psychiatric concern. Of the biographical items initially sought, some were not adequately reported in a sufficient number of cases, and some were not reported in a form amenable to statistical analysis. The following items were ultimately recorded for analysis: school performance; IQ; premorbid social problems manifested by *S*; parental social problems; persons with whom *S* was residing; parental attitude toward having *S*'s problem treated; parental age when *S* was born; parental occupation and education; *S*'s age; number of siblings; birth order; hometown population; and religion.

The data samples of 300 male and 300 female case histories were obtained from the University of Minnesota Hospital Child Psychiatry Unit. As it is a teaching hospital, University Hospital's records are generally more complete than might be the case in nonteaching institutions. Cases were obtained by working backward chronologically from 1964. Certain restrictions were employed in order to insure samples relatively homogeneous with respect to extraneous variables which might influence the clustering of functional symptoms. The following criteria were used to exclude cases from the samples: full scale IQ less than 75; good evidence for organic involvement; serious physical illness or severe chronic physical handicap (e. g., deaf, blind, crippled, spastic); the presence of less than three recordable symptoms in the record; age below 4 years or above 16 years at first psychiatric contact; race other than Caucasian; institutionalization for more than 2 years; residence in a foster home for more than 2 years unless the foster parents were close relatives of the biological parents or unless adopted by the foster parents soon after birth. In addition to these criteria for the exclusion of cases, there was some selection in order to obtain a relatively symmetrical social class distribution. This required going back 1 to 2 years earlier than would otherwise have been necessary in order to secure more upper social class cases. To obtain a median split on age which would be meaningful for certain subgroup analyses, equal numbers of *Ss* in the 4–10 and 11–15 year age groups were sought. For the boys, this required no selective sampling since the median of the sample initially obtained was 10 years. The low frequency of young girls necessitated some selectivity to obtain more of them, and it was possible only to obtain a distribution with a median age of 11. Also, cases having minimal background information were excluded when they could be replaced by more complete cases which met the other criteria. Both inpatients and outpatients were included in the samples. There was no clear-cut distinction between inpatients and outpatients since inpatients were not hospitalized for long periods, and many outpatients eventually became inpatients and vice versa.

The influence of drugs on the symptoms recorded from the present samples was probably minimal. There were no reports of drugs being administered in most of the outpatient functional cases, although it is possible that some of these received tranquilizers through physicians outside the hospital. The use of parent and teacher reports in the data collection was likely to have insured that predrug symptoms were recorded. Cases receiving drugs through hospital physicians usually presented some evidence for organic involvement and would have been excluded from the samples for that reason. A few other cases received drugs while they were inpatients, but this would not have influenced the symptoms recorded from the records, since symptoms occurring only during inpatient treatment were excluded.

The raters were a female college graduate with no specialized training in psychology, and the author who, during the data collection period, was a first- and second-year graduate student in personality research. Approximately 10 cases were initially rated by both raters and disagreements were discussed. Thereafter, each case was read by only one of the raters, except for the 25 randomly selected cases upon which reliability coefficients were calculated. The following instructions were provided for the recording of symptoms:

> Insofar as possible, these (symptom items) refer to behavior and personal reports which require little or no inference. The intake interview, interviews with the parents, letters of referral from doctors, schools, courts, and welfare agencies, and the case summary are the best sources. Those symptoms which have been manifested at some time during the last three years are to be checked if they appear to be in some way a part of the reason for which the child is being referred. For a child of four, enuresis which ceased at age two and one-half and has not been a problem since then would not be included, for example.
> *Caution:* 1. Do *not* check items which appear in the record merely as inferences from psychological instruments or by the interviewer; e. g., at some point in the evaluation of almost every child, the inference will be made that the child is fearful, depressed or the like, but do not

check these items unless, (*a*) the child reports that he is experiencing these feelings; *or* (*b*) there is repeated mention in the record that different people have observed these symptoms; *or* (*c*) there is clear behavioral evidence for these symptoms reported by parents, teachers, or others who have observed the child outside of an interview setting.

2. Do not include items which occur only during the course of inpatient treatment but which were never present before the child was admitted. Do not include common items, e. g. headaches, which are referred to only in the course of the physical exam. Include such items only when they are of abnormal proportions or are also mentioned outside of the physical examination. The initial and final sets of therapy and supervisory staff notes should be read.

3. Do not check more than one symptom on the list for any given item of behavior. For example, if *S* reports having headaches, just the item "headaches" should be checked, while "pains, physical complaints" should not be checked unless there is mention of other physical complaints which are not covered specifically by another item like "stomachaches." Likewise, if the *S* has a strong fear of some specific thing, e. g., dog phobia, the item "phobias, fears" should be checked, but "fearful, anxious" should not be checked unless it is stated that the child is also fearful or anxious in a general non-specific way. If physical causes are found for a symptom, do not include the symptom, e. g., if it is found that blurred vision is being caused by poor eyes.

Each item on the symptom check list is to be regarded as the description of a class of behavior not entirely normal in degree. If behavior fitting one of these class descriptions is noted in the record, that class should be checked, unless the behavior is of apparently normal degree. For example, "fighting" should not be checked for a single mention of "fights with brother," but should be checked if it is frequently mentioned, if it appears to be of abnormal degree, or if it is one of the reasons for which the child was brought to U. M. H.; "crying" should not be checked unless the child cries very easily or is subject to unusual crying spells.

While the use of data from several sources of observations in a record, extracted by raters assumed not to share classical clinical stereotypes and using the methods described above, by no means precludes all sources of bias, the biases remaining are unlikely to be as systematic as the ones possibly influencing the groupings found in previous studies. It must, however, be acknowledged that the attempts at improvement on previous studies fall far short of ideal solutions: larger subject populations would have made still more homogeneous groupings possible, case histories were not always uniform in completeness and clarity, the symptoms sought may not have been properly representative of the symptom domain, and the implementation of the definition of "symptom" may have been subject to unknown influences.

The degree of agreement between the two raters on the 25-case reliability check sample can be calculated in several different ways. The most straightforward method is to calculate the ratio of the number of symptoms on which there was agreement to the total number of symptoms rated. If the number of zero entry symptoms is included in this ratio, that is, the symptoms which were agreed by both raters to be absent, the average percentage of agreement for 25 cases was 96.5 and the median was 96.7, both of which are probably unrealistically high since there was a large proportion of zero entries. If zero entries are not included in the ratio, the average percentage of agreement was 65.5 and the median was 71.4, which may be unrealistically low since zero entries did not necessarily indicate that no judgments were involved. Judgments would have been involved in zero entries where both raters decided that a certain reported observation did not deviate sufficiently from normal to warrant checking the symptom category representing it. Another reliability formula occasionally used in such situations (e. g., Chittenden, 1942; Marshall & McCandless, 1957) is: 2 × sum agreements/# checked by A + # checked by B. Omitting zero entries, this formula yielded an average of 79.1% agreement.

Factor Analyses

The data obtained from the case histories were coded and punched on IBM cards. The data for each sex group were analyzed separately throughout. Symptoms were coded 0 if not checked and 1 if checked. All symptoms which occurred five or more times in a sex group were retained in the analyses for that sex group. Symptoms were intercorrelated using the product-moment correlation routine of the University of Minnesota Statistical Library Program 55 (UMSTAT 55) for orthogonal factor analysis. The symptom intercorrelations are thus represented by phi coefficients. UMSTAT 55 was then used to obtain a principal-factor solution for the correlation matrix. Ones were used in the principal diagonals in place of reliability coefficients, and an eigenvalue of 1.000 (Kaiser's criterion) was set as the minimum below which no factors would be calculated. The limit on the orthogonal rotation angle was set at .009 radians. The entire principal-factor matrix was rotated to the quartimax and varimax approximations to simple structure, using UMSTAT 55. Finally, the first 15 principal factors, that is, those having the 15 largest eigenvalues, were rotated using Carroll's oblimin method, adapted to the 1604 computer by Conrad Katzenmeyer. The gamma value for the oblimin rotation was set at .5. The fact that the oblimin factors can be intercorrelated means that the addition of each new principal factor can cause marked changes in the rotated factors obtained before the addition of that principal factor. Because the program printed out the solution each time another principal factor was added, 14 oblimin solutions were obtained.

Although simple structure is the predominant rotational standard for factor analysis, its cri-

teria are not mathematically precise. The use of two different orthogonal rotations and 14 oblimin rotations was intended to aid in the identification of especially reliable factors. It was expected that a factor which appeared in each of the various rotations would be a relatively reliable one since changes in rotational criteria did not obscure it.

Another objective of the factor-analytic approach employed was to seek out hierarchical orderings of groupings which might be present in the data. The most obvious means to this end was through the second-order analysis of the oblique factors obtained in the oblimin rotations.

RESULTS

MALES

The Sample

The distribution of the samples by age, parental social class, and diagnosis can be obtained from the tables in Appendix B. The chronological span of admission dates covered in this sample was 1953 through 1964. As a result of the effort to exclude individuals who had been adopted later than early infancy, only five adopted cases appeared in the sample, the oldest age at adoption being 3½ months. The exclusion of individuals institutionalized over long periods resulted in the inclusion of only four cases which could in any sense be considered institutionalized prior to psychiatric referral. Two of these cases had been legally committed to state-aided foster homes because of their difficult behavior, one was in reform school, and one in a detention center. In none of these cases did the period of "institutionalization" exceed 5 months. Thus the vast majority of the boys were living at home with at least one biological parent or with close relatives.

After excluding all symptoms which occurred less than five times in the sample of 300 cases, 74 symptoms remained for analysis. These are indicated by the letter M on the symptom check list in Appendix A. Counting only those 74 symptoms retained for analysis, the mean number of symptoms recorded per case was 8.28.

Factor Analyses

In this section, the factors presented are given discriptive names for convenience of reference.[4] These names were selected through consultation with one adult and three child clinical psychologists. The names are not intended to convey theoretical or interpretive implications but are merely attempts at shorthand descriptions which will be meaningful to psychologists. In some instances it was difficult to get consensus on a descriptive label, and several qualifications accompany the label finally selected. All minus signs are dropped from the factor loadings presented. The items are listed as shortened versions of those used in data collection, but the full symptom category, as listed in Appendix A, is always implied. For example, "suicidal" on a factor represents *Item 45*, "masochism, self-harm, suicidal, threatens to kill self."

Principal factors. Five principal-factor analyses were performed on the symptom data of the male sample. First, the symptoms from the entire group of 300 were intercorrelated and analyzed. Second, a median split on age produced two groups of 150 *S*s each, ranging in age from 4–10 and 11–15, respectively. In each of these groups, symptoms occurring four or more times (62 for the younger males and 67 for the older males) were intercorrelated and analyzed. Third, a median split on social class produced a group of 144 *S*s in the lower three classes and 155 *S*s in the upper three classes, with one *S* being excluded due to lack of social class data. Symptoms occurring four or more times (65 in the lower class group and 66 in the upper class group) were intercorrelated and analyzed separately for each group. Age and social class were evidentally independent in these dichotomies since approximately equal numbers of lower and upper class *S*s fell into the younger and older groups. The median splits were rather coarse breakdowns, but meaningful factor analyses would have been precluded by the small *N*s resulting from finer breakdowns.

[4] In a previous brief summary of this research (Achenbach, 1965), some of the factors were given labels slightly different from those reported here. In addition, a few small rotated factors which classified very few *S*s, the second principal factors for both sexes, and two rotated factors for the girls which did classify a significant number of *S*s had not been fully investigated by the time of that report. Except for the addition of these factors, the overall pattern of classification of *S*s there did not differ from the one presented here.

TABLE 1
MALE PRINCIPAL FACTORS

(a) First principal factor

Internalizing (positive end of first principal factor)	Externalizing (negative end of first principal factor)
.526 Phobias	.632 Disobedient
.424 Stomachaches	.555 Stealing
.382 Fearful	.510 Lying
.363 Pains	.492 Fighting
.344 Worrying	.453 Cruelty
.343 Withdrawn	.445 Destructive
.339 Nausea	.399 Inadequate guilt feelings
.335 Obsessions	.398 Vandalism
.330 Shy	.372 Truancy
.329 Vomiting	.362 Fire-setting
.304 Compulsions	.342 Swearing
.302 Insomnia	.317 Running away
.266 Crying	.277 Temper tantrums
.263 Fantastic thinking	.275 Showing off
.259 Headaches	.274 Hyperactive
.256 Seclusive	.227 Sexual delinquency
.249 Apathy	.201 Threatening people
.231 Depression	.149 Negativistic
.227 Nightmares	.144 Poor school work
.225 Nervous	.133 Sexual perversions
.220 Refusing to eat	.121 Attention demanding
.185 Overtired	.108 Enuresis
.182 Fears own impulses	.103 Encopresis
.154 Confused	
.153 Self-conscious	
.151 Obese	
.135 Tics	
.120 Bizarre behavior	
.117 Stuttering	
.107 Skin eruptions	
.101 Asthma	

(b) Second principal factor

Severe and Diffuse Psychopathology (unipolar)

.531 Bizarre behavior	.253 Swearing
.491 Fantastic thinking	.246 Phobias
.460 Temper tantrums	.246 Attention demanding
.418 Threatening people	.229 Moodiness
.379 Ideas of reference	.226 Can't concentrate
.372 Insomnia	.222 Vomiting
.359 Loudness	.215 Quarrelsome
.326 Nightmares	.215 Withdrawn
.320 Crying	.212 Headaches
.268 Disobedient	.212 Daydreaming
.268 Nausea	.208 Poor motor coordination
.267 Destructive	.207 Obsessions
.260 Fighting	.207 Refusing to eat
.257 Cruelty	

The number of principal factors obtained in the five analyses ranged from 23 to 28. In each analysis, the first principal factor was bipolar, and its eigenvalue was substantially larger than those of the second and succeeding factors. The pattern of symptom loadings on the first principal factor was also very similar throughout the five analyses. Congruence coefficients, calculated by the product-moment correlation formula (Harmon, 1960), between the first principal factor from the analysis of the entire sample and the first principal factors from each of the four subgroups ranged from .956 to .985, with a mean of .968. This indicated a highly reliable dimension which was similar in each of the relatively homogeneous subgroups and of which the first principal factor for the entire group was a very good representative. Even relatively low factor loadings appeared to reflect the polar tendencies of this principal factor, and items with loadings as small as $\pm.100$ are presented in Table 1. The label Internalizing versus Externalizing was selected for this factor. The label is not intended to carry dynamic implications. It means only that the symptoms at the Externalizing end describe conflict with the environment, while those at the other end describe problems within the self.

Although having a substantially smaller eigenvalue than the first principal factor (3.080 versus 5.269 in the analysis of the entire sample), the second principal factor appeared relatively consistent throughout the five analyses. It was unipolar, and congruence coefficients between the second principal factor from the analysis of the entire sample and those from the four subgroups ranged from .722 to .911, with a mean of .790. Finding a descriptive label which commanded consensus was much more difficult for this factor than for the first one. Suggested labels included "severe psychopathology," "extreme disorganization," and "diffuse psychopathological disorganization." Descriptively, it seemed to include symptoms both of severe mental disturbance and of excessive belligerence. The label Severe and Diffuse Psychopathology was finally chosen. Items with loadings down to .200 are presented in Table 1.

Rotated factors. Because the largest two factors were so similar in the five analyses, and the succeeding factors had relatively low eigenvalues (2.684 for the third factor, 2.282 for the fourth in the analysis of the entire sample), the primary dimensions in

the principal-factor matrix from the entire sample were assumed to be representative of the various subgroups, and this matrix was rotated to obtain the different simple structure solutions. All 28 principal factors were rotated to the varimax and quartimax criteria, and the 15 principal factors having the largest eigenvalues were rotated to the oblimin criterion. The resulting approximations to simple structure were examined for factors which were identifiable in all or most of the solutions. Any pattern of symptom loadings which appeared consistently on a factor throughout the rotated solutions was concluded to represent a relatively reliable factor. All variations of such a pattern of symptom loadings were then considered together and ranked according to how representative of the group they seemed to be. No items with loadings below .100 were considered in the selection procedure. A subjective combination of the following criteria was used in choosing the best representative for each factor: (*a*) inclusion of the most symptoms which frequently appeared on the other variations of the factor; (*b*) least overlap of symptoms with other reliable factors; (*c*) highest factor loadings. In addition, for each factor, a cutoff point was selected for loadings below which items would not be considered because they began to overlap with another factor. The following example may help to illustrate this procedure: it was noticed that a factor on which "vomiting" and "nausea" had the highest loadings, and which usually included "stomachaches," "headaches," "pains," and "phobias," appeared in 15 of the 16 rotated solutions (varimax, quartimax, and 13 of the 14 oblimin solutions). Each of the 15 variations of this factor, including all symptoms with loadings of .100 and above, was typed on a separate card. Then the symptoms with the lowest loadings on each variation of the factor were examined to see if there was a point where groups of symptoms consistently having high loadings on another group of rotated factors appeared together. It was found that "fantastic thinking," "confused," and "withdrawn," which appeared together with relatively high loadings on another group of rotated factors, appeared together with low loadings on most of the variations of the "vomiting-nausea" factor. Therefore, "fantastic thinking" and items with lower loadings than "fantastic thinking" were dropped from further consideration in relation to this factor. The remaining symptoms on each of the 15 variations were then used to rank the 15 according to the three criteria stated above. The third factor in the three-factor oblimin solution was chosen by this method as the best representative of the group of factors heavily loaded on "vomiting and nausea." The four clinical psychologists who were later consulted for labeling the factors took into consideration the five variations which had been ranked as the five best representatives of each factor.

By the foregoing procedure eight rotated factors were found. These are presented in Table 2, and the rotated solution in which it appeared is indicated for each one. Complete consensus was attained in labeling four of them: 1. Somatic Complaints; 2. Delinquent Behavior; 3. Obsessions, Compulsions, and Phobias; and, 4. Sexual Problems. Factor 5 was generally agreed to include schizoid thinking and behavior, but some reservation was voiced due to the possible implication that "schizoid" might imply relatively mild pathology while the symptoms on the factor included some that might be severe enough to be called schizophrenic. Schizoid Thinking and Behavior was selected with the qualification that nothing about intensity of behavioral deviation is implied. Neither is any contrast with "schizophrenic" nor any implication about etiology intended. "Schizoid" would simply appear to be a more general term than "schizophrenic."

Factors 6 and 7 presented greater labeling difficulties. Factor 6 elicited suggestions of "unsocialized aggressive behavior," "antisocial aggression," and "aggressive." The common element in these suggestions was aggression, so the label Aggressive Behavior was chosen, with the qualification that some of the behavior included was not direct aggression against other persons.

Factor 7 elicited suggestions of "uncontrolled impulsivity and hyperactivity," "hyperactive agitation," "tenseness, hyper-

TABLE 2

MALE ROTATED FACTORS

1. Somatic Complaints (oblimin 3-3)
   - .534 Vomiting
   - .525 Nausea
   - .517 Stomachaches
   - .501 Headaches
   - .418 Pains
   - .374 Phobias
   - .371 Depression
   - .299 Suicidal
   - .297 Overtired
   - .295 Ideas of reference
   - .283 Dizziness
   - .276 Insomnia
   - .263 Withdrawn
   - .256 Diplopia
   - .207 Worrying
   - .191 Nightmares
   - .183 Shy
   - .164 Obese
   - .162 Complains no one loves him
   - .156 Fearful
2. Delinquent Behavior (varimax 1)
   - .719 Truancy
   - .697 Running away
   - .509 Vandalism
   - .493 Lying
   - .479 Inadequate guilt feelings
   - .476 Disobedient
   - .327 Fire-setting
   - .243 Fighting
   - .238 Cruelty
   - .196 Swearing
   - .179 Sexual delinquency
   - .162 Destructive
   - .132 Poor school work
   - .102 Showing off
3. Obsessions, Compulsions, and Phobias (oblimin 4-2)
   - .647 Obsessions
   - .545 Compulsions
   - .465 Fears own impulses
   - .448 Fearful
   - .390 Tics
   - .350 Stuttering
   - .329 Phobias
   - .318 Nervous
   - .309 Seclusive
   - .306 Fantastic thinking
   - .295 Self-conscious
   - .294 Daydreaming
   - .287 Insomnia
   - .258 Worrying
   - .243 Withdrawn
   - .217 Nightmares
4. Sexual Problems (oblimin 15-4)
   - .607 Masturbation
   - .500 Sexual preoccupation
   - .451 Sexual delinquency
   - .353 Overtired
   - .297 Thumbsucking
   - .217 Swearing
   - .174 Lying
   - .173 Enuresis
   - .165 Withdrawn
   - .163 Sexual perversions
5. Schizoid Thinking and Behavior (quartimax 2)
   - .745 Fantastic thinking
   - .577 Bizarre behavior
   - .461 Insomnia
   - .428 Loudness
   - .385 Crying
   - .316 Confused
   - .267 Phobias
   - .231 Poor motor coordination
   - .215 Withdrawn
   - .203 Refusing to eat
   - .202 Ideas of reference
   - .198 Overtired
   - .189 Quarrelsome
   - .187 Stuttering
   - .186 Nightmares
   - .177 Obsessions
   - .172 Temper tantrums
   - .170 Headaches
   - .146 Daydreaming
   - .134 Fearful
   - .114 Can't concentrate
6. Aggressive Behavior (quartimax 14)
   - .813 Cruelty
   - .554 Threatening people
   - .546 Destructive
   - .402 Inadequate guilt feelings
   - .373 Vandalism
   - .300 Disobedient
   - .215 Suicidal
   - .194 Lying
   - .183 Ideas of reference
   - .157 Temper tantrums
   - .139 Fire-setting
   - .134 Stealing
   - .124 Truancy
   - .124 Fighting
7. Hyperreactive Behavior (oblimin 3-3, negative end)
   - .494 Hyperactive
   - .360 Can't concentrate
   - .282 Attention demanding
   - .216 Loudness
   - .209 Quarrelsome
   - .182 Tics
   - .159 Stuttering
   - .158 Encopresis
   - .129 Fighting
   - .115 Poor school work
   - .102 Temper tantrums
8. (Unnamed) (oblimin 11-11)
   - .595 Constipation
   - .463 Nailbiting
   - .449 Encopresis

TABLE 2—*Continued*

.401 Dizziness	.164 Poor motor control
.352 Destructive	.160 Shy
.340 Enuresis	.153 Skin eruptions
.194 Fire-setting	.147 Stomachaches
.173 Diplopia	.128 Insomnia
.173 Fighting	.126 Overtired

Note.—The source of each factor is indicated in parentheses: e.g., (oblimin 4-2) means the factor was the second factor in the four factor oblimin solution; (varimax 1) means it was the first factor in the varimax solution.

activity, and irritability," and "hyperactive-hypersensitive." Hyperactivity was the obvious common element in all these suggestions, and a common reference to overreactivity to stimulation was also made. With reservations, the term Hyperreactive Behavior was finally chosen as being general enough to include both the items suggestive of hyperactivity and those suggestive of impulsivity-agitation-irritability-hypersensitivity. Factor 8 was found to classify, by means of the procedure to be described below, only two *S*s out of the 300, so it was left unnamed and was not retained for further statistical analyses.

Second-order factors. The goal of obtaining higher order groupings of symptoms was to be attained by means of orthogonal factor analyses of the intercorrelations between first-order oblimin factors. An initial difficulty, however, was that a choice of oblimin solution had to be made since there were 14 solutions available. Because not all the oblimin factors chosen as being the best representatives of their respective groups came from the same solution, no single solution could be chosen by that particular criterion. It was decided, therefore, to do second-order analyses of two oblimin solutions which were not highly similar but which contained good examples of most of the reliable factors. The four- and eight-factor oblimin solutions were chosen, and the factor intercorrelations were analyzed by the principal-factor method with quartimax and varimax rotations. The principal-factor solution was similar to the rotated solutions in both of these analyses. In the second-order analysis of the four-factor oblimin solution, two factors appeared. In all three second-order orthogonal solutions, the first factor was bipolar, with high loadings on the Somatic Complaints and Obsessions, Compulsions, and Phobias factors at one end and a high loading on the Delinquent Behavior factor at the other end. The second factor in each solution was unipolar, with the highest loading on the Schizoid Thinking and Behavior factor. The three second-order solutions for the eight oblimin factors were also very similar. They produced three factors, the first of which was bipolar with high loadings on the Somatic Complaints and Obsessions, Compulsions, and Phobias factors at one end and a high loading on Sexual Problems at the other end. The third factor tended to be unipolar although the smaller pole had some relatively large loadings. Hyperreactive Behavior had the heaviest loading on this factor.

To obtain loadings for specific symptoms on the second-order factors, the obvious procedure is to multiply the loadings of the symptoms of a first-order factor by the loading that factor has on a second-order factor. Drawing conclusions from the results of this procedure is difficult because any symptoms which have loadings on more than one first-order factor may be given several different second-order loadings. The situation is further complicated when some of these second-order loadings are negative and some positive, which may occur if a symptom appears at the positive end of one first-order factor having a positive loading on a second-order factor and the same symptom appears at the negative end of another first-order factor also having positive loadings on the same second-order factor. A similar situation may arise if a symptom has positive loadings on two first-order factors, one of which is loaded positively and the other negatively on a given second-order factor. Empirically, these possibili-

ties may be somewhat remote among symptoms with high first-order loadings, but they did occasionally arise for symptoms with smaller loadings. To simplify the results somewhat, only positive first-order loadings were multiplied by the second-order loadings since the oblimin factors generally tended toward unipolarity with their heaviest loadings on the positive end. Also, only first-order factors with second-order loadings of .300 or above were included in the calculations.

There was an obvious similarity between the bipolar groupings on the first-order first principal factor (Table 1) and the bipolar groupings on the largest second-order factor for both the four- and eight-factor oblimin solutions. There is good reason for expecting this similarity between the first principal factor and the first unrotated second-order factor. Rimoldi (1951; cf. Fruchter, 1954, p. 172) subjected the same intercorrelation matrix of reasoning tests both to multiple-group and to Spearman type two-factor analyses. The intercorrelations of the factors obtained by the multiple-group method were analyzed to three centroid factors. The loadings of the tests on the first unrotated second-order factor were found to be approximately proportional to their loadings on the *g* factor produced by the two-factor method. Since the first principal factor in a principal-factor solution is similar to a Spearman *g* factor (Burt, 1938; cf. Harmon, 1960, p. 163), and since the multiple-group method applied to variables with well-known properties, such as the tests used by Rimoldi, can resemble oblique rotations to simple structure, it is not surprising that the first principal factor and first unrotated second-order factor found in the present study were similar in their patterns of item loadings. It would seem intuitively probable that a *g* type first-order factor should be similar to the largest second-order factor if they both accurately reflect the most central dimension in the correlation matrix. In any event, for practical purposes, the first principal factor appearing in the present data can probably be regarded as representing the primary dimension in the data better than do any of the possible second-order factors for the following reasons: (*a*) as a least squares solution, the first principal factor is mathematically unique whereas the oblimin solution on which second-order factors are based is a mathematical approximation to the criteria for simple structure which themselves do not determine a mathematically unique solution; (*b*) there are many oblimin solutions from which the second-order factors can be obtained and there are no precise criteria for selecting one as the most appropriate; (*c*) multiplying the first-order item loadings by the second-order factor loadings yields more than one value for some items and may yield positive and negative second-order loadings for the same item; and, (*d*) the first principal factor here did not suffer from weaknesses usually attributed to unrotated factors as compared to rotated ones—specifically, lack of meaningfulness and lack of stability when items are dropped; the first principal factor reflected a dichotomy which was descriptively consistent and psychologically meaningful without rotation, and, as evidenced by the high congruence coefficients reported above, it was barely altered by dropping as many as 12 items and by analyzing different subgroups of *S*s.

Classification of Ss *by the factors*. To determine what meaning the factors had for the classification of individual cases and what the relationship was between the general Internalizing-Externalizing dimension and the rotated factors, cases were classified according to their standing on the various factors reported in Tables 1 and 2. The classification of cases solely on the basis of factor scores was deemed inappropriate since each case had many zero entry items (symptoms not reported as present). This would have caused factor scores to have been badly confounded with the number of symptoms reported, which in turn may occasionally have been a function of the completeness of the case record. Furthermore, since factors were taken from different solutions, their loadings were not necessarily comparable on a single absolute scale.

To avoid these two sources of error, it was decided to classify *S*s according to an arbitrarily selected degree of resemblance

to a factor, in terms of the percentage of their symptoms matching that factor. First, if 60% or more of an *S*'s symptoms came from the internalizing end of the first principal factor, he was assigned to the Internalizing category. Likewise, if 60% or more of his symptoms came from the externalizing end, he was assigned to the Externalizing category. By this criterion, 68 cases were assigned to the Internalizing category, 128 cases were assigned to the Externalizing category, and 104 cases were left unclassified. The large number of unclassified cases is not simply a result of almost equal proportions of Internalizing and Externalizing symptoms occurring in all these cases. Only the 54 symptoms with loadings of ±.100 were used in classifying cases, but the 20 symptoms having loadings below ±.100 also contributed to the denominator of the proportion calculated for each *S*. Cases whose symptoms included some of these 20 might therefore fail to meet the 60% criterion, even though the numbers of Internalizing and Externalizing symptoms were far from equal. Several other means of classification could also have been attempted. For example, a continuous scoring based on the precise percentage of symptoms from each group could have been employed. The precision of the data was not thought to warrant such scoring however. Future ap-

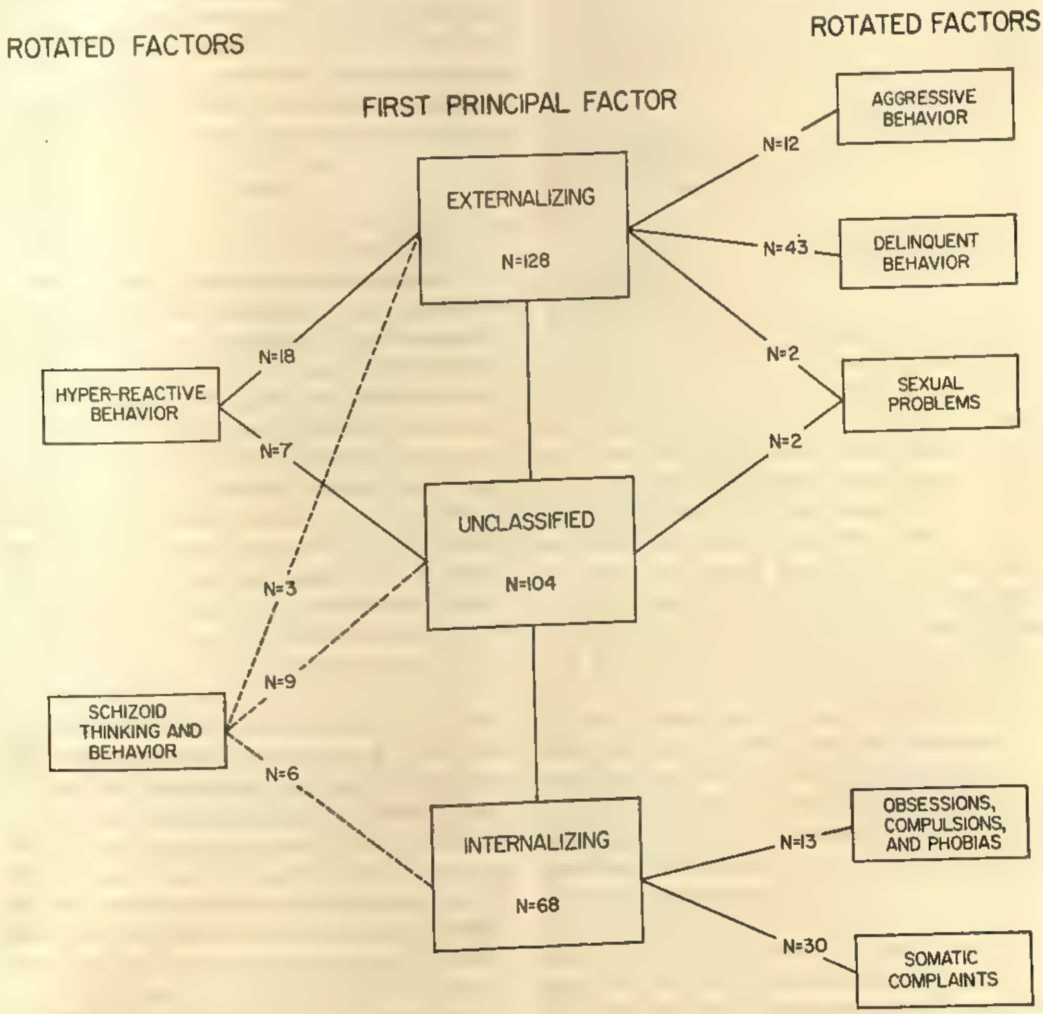

Fig. 1. Classification of male *Ss* by first principal and rotated factors. (One "obsessive" and two "somatic" *Ss* came from the Unclassified group.)

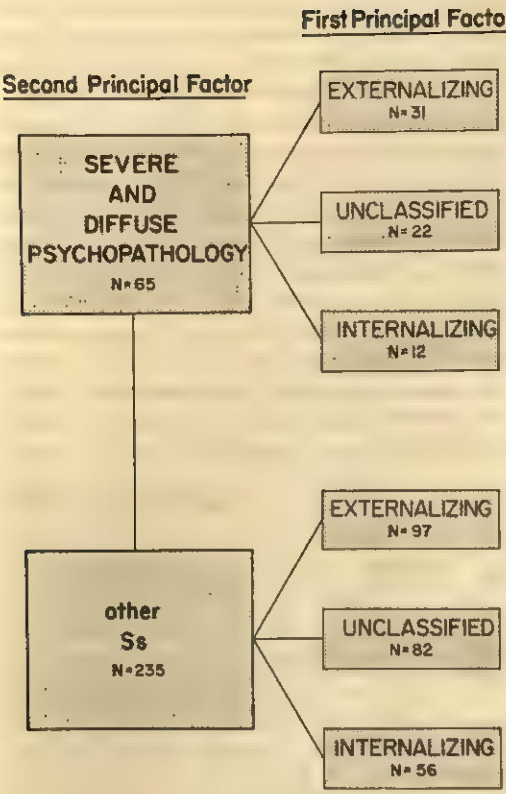

FIG. 2. Classification of male *Ss* by first and second principal factors.

plications of the present factors to live patients should seek quantitative refinements.

The second step was to classify cases by their resemblance to the rotated factors. Again, if 60% or more of an *S*'s symptoms appeared on a rotated factor, he was assigned to the category represented by that factor. However, unlike the mutually exclusive polar groupings of symptoms on the principal factor, the groupings of symptoms on the rotated factors overlapped, raising the possibility of tied scores. If an *S* had the same percentage of symptoms (at or above the 60% criterion) on two or more rotated factors, factor scores were calculated, and the *S* was assigned the category for which his factor score was highest. Figure 1 presents the number of *Ss* classified by the principal factor, by each of the rotated factors, and the relationship between the two levels of classification.

Since the second principal factor also showed some consistency in the several subgroup analyses and gave some evidence of appearing in the second-order analyses, it too was used to classify cases by the 60% criterion. Because it was a unipolar factor and generally showed less stability than the first principal factor, only symptoms with loadings of .200 and above, as presented in Table 1, were used for the classification procedure. Sixty-five cases were found to meet the 60% criterion, while the other 235 did not. A comparison of the classification of cases by the first principal factor and by the second principal factor (Figure 2) shows that they were clearly orthogonal to one another. The cases falling into the category of Severe and Diffuse Psychopathology were distributed in the Internalizing, Externalizing, and Unclassified categories of the first principal factor roughly proportionally to the total number of cases in those categories.

FEMALES

The Sample

The chronological span of admission dates was 1951 through 1964. Nine of the *Ss* had been adopted, the oldest age at adoption being 7 months. Only two of the cases could be considered institutionalized prior to referral. Both of these had been living in state-supported boarding homes, one for 2 months and one for 1 year. Seventy-three symptoms occurred five or more times, and these are indicated by the letter F on the symptom check list in Appendix A. Counting only these 73 symptoms, the mean number of symptoms recorded per case was 7.69.

Factor Analyses

Exactly the same factor analytic procedures as with the males were followed. The median split on age resulted in two groups of 150 *Ss* each, ranging in age from 4–11 and 12–15. In the younger group 60 symptoms occurred four or more times, and in the older group 65 symptoms occurred four or more times. The median split on social class produced a group of 160 *Ss* in the lower three classes and 130 in the upper three classes, with social class data being unavailable for 10 *Ss*. Sixty-six symptoms in the lower class group and 64 symptoms in

the upper class group occurred with a frequency of four or greater. Approximately equal numbers of lower class and of upper class *S*s appeared in each age group, indicating that the social class and age dichotomies were independent.

Principal factors. The number of principal factors obtained in the five analyses ranged from 24 to 28. As with the males, the eigenvalue of the first principal factor was substantially larger than that of the second factor, the pattern of symptom loadings was similar throughout the five analyses, and the bipolar distribution of symptoms represented a dichotomy similar to that found for the males. Congruence coefficients between the first principal factor from the analysis of the entire sample and the first principal factor from each of the four subgroups ranged from .848 to .967, with a mean of .922. The factor has been given the same name as that for the males and is presented in Table 3.

The second principal factor was less consistent than in the male analyses. Like the males' factor, it tended to be unipolar and difficult to label descriptively. It appeared to include many of the same extremely pathological items as the males' factor, but without the more aggressive behaviors. Severe and Diffuse Psychopathology seemed to apply as well to this factor as to the males' factor, although it was clear that this factor contained more symptoms found at the Internalizing pole of the first principal factor than did the male factor, which included more symptoms from the Externalizing pole of the male first principal factor. The congruence coefficients between the second principal factor from the entire sample and those from the subgroups were relatively low, ranging from .110 to .538, with a mean of .420. Three of these coefficients were in the .50's however. The factor which had a congruence coefficient of only .110 appeared in the analysis of the upper social class female data. It tended to be bipolar with symptoms similar to those heavily loaded on the other second principal factors at one end and a few symptoms like "nausea," "stomachaches," and "inappropriately indifferent" heavily loaded at the other end. Despite the low congruence coefficient, 14 out of the 22 symptoms with loadings above .200 on the larger pole were also loaded above .200 on the pole of the factor from the entire sample. The factor from the entire

TABLE 3

FEMALE PRINCIPAL FACTORS

(a) First principal factor

Internalizing (negative end of first principal factor)	Externalizing (positive end of first principal factor)
.521 Nausea	.562 Disobedient
.494 Pains	.512 Lying
.483 Headaches	.447 Stealing
.448 Stomachaches	.387 Fighting
.426 Phobias	.323 Running away
.346 Vomiting	.320 Swearing
.324 Diplopia	.317 Quarrelsome
.295 Refusing to eat	.281 Threatening people
.290 Obsessions	.267 Truancy
.285 Fearful	.249 Destructive
.281 Withdrawn	.227 Poor school work
.274 Depression	.224 Attention demanding
.264 Dizziness	.224 Sexual delinquency
.259 Crying	.212 Inadequate guilt feelings
.242 Nightmares	.184 Sexual preoccupation
.239 Nervous	.176 Thumbsucking
.233 Worrying	.174 Masturbation
.230 Insomnia	.165 Enuresis
.228 Constipation	.155 Temper tantrums
.223 Fears own impulses	.148 Negativistic
.221 Breathing difficulty	.147 Nailbiting
.210 Compulsions	.115 Hyperactive
.209 Shy	.106 Poor motor coordination
.183 Overtired	
.166 Self-conscious	
.161 Confused	
.150 Fantastic thinking	
.149 Inappropriately indifferent	
.142 Tics	
.127 Skin eruptions	
.115 Feelings of worthlessness	
.111 Obese	

(b) Second principal factor

Severe and Diffuse Psychopathology (unipolar)

.478 Withdrawn	.310 Temper tantrums
.406 Bizarre behavior	.306 Fantastic thinking
.388 Confused	.296 Obsessions
.381 Depression	.288 Negativistic
.366 Ideas of reference	.270 Moodiness
.352 Crying	.265 Destructive
.338 Fearful	.262 Hyperactive
.330 Fears own impulses	.233 Seclusive
.328 Compulsions	.233 Feelings of worthlessness
.328 Can't concentrate	.229 Phobias
.324 Insomnia	.221 Excessive talking
.313 Daydreaming	

TABLE 4

FEMALE ROTATED FACTORS

1. Somatic Complaints (oblimin 5-3)
   - .674 Headaches
   - .599 Stomachaches
   - .597 Nausea
   - .566 Pains
   - .493 Vomiting
   - .493 Dizziness
   - .450 Breathing difficulty
   - .423 Diploplia
   - .307 Inappropriately indifferent
   - .305 Overtired
   - .297 Fainting
   - .173 Tics
   - .152 Refusing to eat
   - .150 Constipation
   - .144 Phobias
   - .135 Nervous
2. Delinquent Behavior (quartimax 6, negative end)
   - .770 Lying
   - .710 Stealing
   - .391 Inadequate guilt feelings
   - .297 Disobedient
   - .274 Masturbation
   - .237 Attention demanding
   - .222 Nailbiting
   - .220 Truancy
   - .136 Destructive
   - .133 Fighting
   - .112 Temper tantrums
3. Obsessions, Compulsions, and Phobias (oblimin 7-2)
   - .590 Fears own impulses
   - .577 Obsessions
   - .546 Compulsions
   - .453 Phobias
   - .390 Crying
   - .365 Refusing to eat
   - .358 Feelings of worthlessness
   - .346 Depression
   - .343 Confused
   - .291 Fearful
   - .280 Insomnia
   - .279 Withdrawn
   - .255 Worrying
   - .225 Moodiness
   - .192 Nightmares
   - .180 Refusing to talk
   - .167 Constipation
4. Schizoid Thinking and Behavior (oblimin 10-8)
   - .662 Fantastic thinking
   - .628 Bizarre behavior
   - .410 Confused
   - .401 Ideas of reference
   - .366 Negativistic
   - .329 Seclusive
   - .255 Apathy
   - .240 Fighting
   - .182 Destructive
   - .179 Fearful
   - .173 Poor school work
   - .172 Headaches
   - .170 Poor motor coordination
   - .162 Withdrawn
   - .150 Daydreaming
   - .146 Insomnia
   - .137 Sexual preoccupation
   - .128 Refusing to talk
   - .124 Constipation
   - .123 Encopresis
   - .120 Obsessions
5. Aggressive Behavior (quartimax 22, negative end)
   - .666 Swearing
   - .625 Threatening people
   - .583 Fighting
   - .555 Destructive
   - .441 Disobedient
   - .312 Temper tantrums
   - .208 Suicidal
   - .179 Truancy
   - .162 Attention demanding
   - .158 Quarrelsome
6. Hyperreactive Behavior (oblimin 4-4)
   - .582 Hyperactive
   - .487 Can't concentrate
   - .343 Nervous
   - .323 Tics
   - .313 Attention demanding
   - .295 Thumbsucking
   - .274 Nailbiting
   - .274 Excessive talking
   - .230 Destructive
7. Depressive Symptoms (oblimin 11-9)
   - .650 Depression
   - .516 Moodiness
   - .458 Withdrawn
   - .447 Suicidal
   - .364 Temper tantrums
   - .279 Crying
   - .262 Refusing to eat
   - .226 Running away
   - .220 Shy
   - .212 Fearful
   - .210 Compulsions
   - .205 Complains no one loves him
   - .184 Self-conscious
   - .149 Daydreaming
   - .140 Breathing difficulty
   - .130 Feelings of worthlessness
   - .129 Refusing to talk
8. Neurotic and Delinquent Behavior (oblimin 9-6)
   - .430 Truancy
   - .362 Poor school work
   - .358 Running away
   - .329 Overtired
   - .322 Sexual delinquency
   - .317 Asthma
   - .304 Disobedient

TABLE 4—*Continued*

.295 Suicidal
.243 Skin eruptions
.207 Depression
.204 Fainting
.150 Breathing difficulty
.135 Poor motor coordination
.134 Daydreaming
.129 Confused
.123 Swearing
.108 Can't concentrate
.105 Apathy
.102 Lying

9. Obesity (oblimin 5-5)
.666 Obese
.629 Self-conscious
.607 Overeating
.392 Shy
.324 Withdrawn
.312 Depression
.263 Daydreaming
.217 Loneliness
.199 Complains no one loves him
.192 Suicidal
.154 Pains
.140 Nausea
.133 Overtired

10. Anxiety Symptoms (oblimin 9-4)
.535 Insomnia
.381 Crying
.359 Phobias
.347 Worrying
.340 Skin eruptions
.325 Nightmares
.297 Asthma
.266 Vomiting
.260 Nausea
.236 Picking
.186 Nervous
.173 Destructive
.137 Thumbsucking
.121 Fearful
.107 Poor motor coordination
.105 Refusing to eat

11. Enuresis and Other Immaturities (oblimin 14-3, negative end)
.490 Enuresis
.291 Thumbsucking
.263 Refusing to eat
.249 Stuttering
.203 Encopresis
.195 Shy
.165 Masturbation
.151 Apathy
.146 Destructive
.129 Refusing to talk
.123 Overeating
.121 Constipation
.109 Temper tantrums

sample is presented in Table 3. The third and succeeding factors had relatively small eigenvalues compared to the first two principal factors (4.479 for the first, 3.321 for the second, 2.647 for the third, and 2.375 for the fourth).

Rotated factors. The same rotational and selection procedures used with the males' data yielded 11 factors which were considered reliable (Table 4). Five of these were agreed by the clinical consultants to resemble reliable factors found for the males, and they were therefore given the same descriptive labels. These were: 1. Somatic Complaints; 2. Delinquent Behavior; 3. Obsessions, Compulsions, and Phobias; 4. Schizoid Thinking and Behavior; and 5. Aggressive Behavior. Factor 6 presented among its more heavily loaded items a pattern somewhat similar to that of the male factor labeled Hyperreactive Behavior and was given the same label, with the qualification that more symptoms from the Internalizing group were present. Factor 7 was agreed to be well described by the label Depressive Symptoms. Factor 8 was agreed to include both symptoms commonly labeled neurotic and behaviors labeled delinquent. Some reservations about the label Neurotic and Delinquent Behavior were expressed, but, since no other label was suggested, this was retained. Obesity was agreed to be the most applicable single descriptive term for Factor 9, although it was suggested that the other items on the factor were also descriptive of social inadequacy and the oral depressive syndrome. Because it was difficult to choose a term to describe this aspect of the factor, the label Obesity was applied with the qualification that the factor included several symptoms and behaviors beside the physical condition of obesity.

There was consensus that Factor 10 was composed mainly of symptoms usually attributed to anxiety, so it was labeled Anxiety Symptoms. The labeling of Factor 11 presented considerable difficulty. It was pointed out that it included several symptoms often listed under Special Symptom Reaction in the Standard Nomenclature. It was also pointed out that the items reflected a common immaturity or lack of control in

regulatory functions. Enuresis was by far the most heavily loaded symptom on the several variations of the factor. It bore some similarity to the unnamed Factor 8 found in the male analyses. Since enuresis was clearly the most prominent item, and there was some agreement as to the immature quality of the other items, the label Enuresis and Other Immaturities was finally selected.

Second-order factors. As was done with the male data, the four- and eight-factor oblimin solutions for the female data were subjected to second-order orthogonal analyses. The two second-order orthogonal factors obtained from the four oblimin factors were very similar in the principal factor, varimax, and quartimax solutions. The two largest second-order factors obtained from the eight oblimin factors were also very similar for the three orthogonal solutions, while the third orthogonal factor was consistent only with respect to the most heavily loaded first-order factors. The largest second-order principal factor for both the four- and eight-factor oblimin solutions was bipolar and resembled the dichotomy found in the first-order first principal factor. The second-order second principal factor was unipolar in both cases and showed some resemblance to the first-order second principal factor.

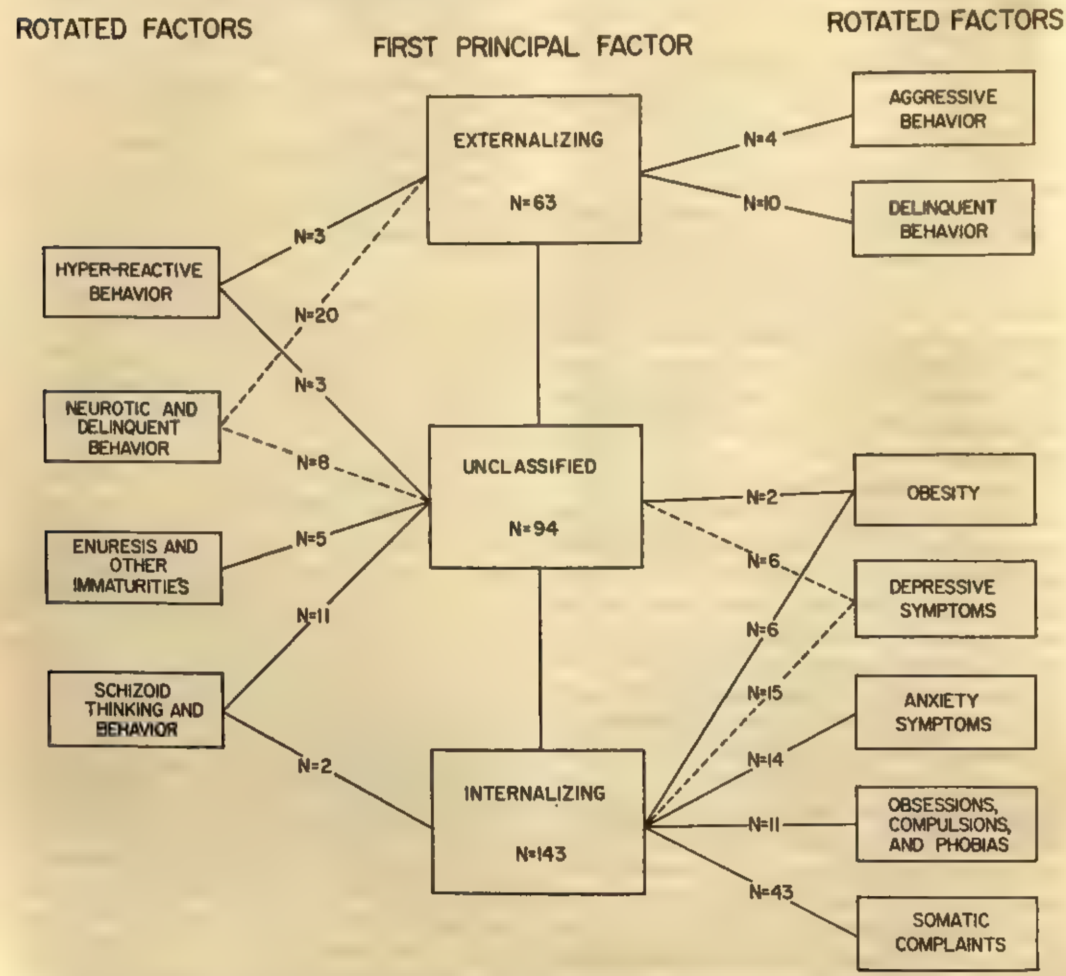

Fig. 3. Classification of female *S*s by first principal and rotated factors. (One "obsessive" and one "anxiety" *S* came from the Unclassified group; one "enuresis" *S* came from the Externalizing group; one "enuresis" and one "neurotic and delinquent" *S* came from the Internalizing group.)

Classification of Ss *by the factors.* For the reasons advanced earlier, it was decided that the first-order first principal factor was the best representative of the most primary dimension in the matrix of symptom intercorrelations. Replication of the classification procedure employed with the males resulted in the assignment of 143 females to the Internalizing category, 63 to the Externalizing category, and 94 to the Unclassified group. Figure 3 presents the number of cases assigned by the same method to each of the categories represented by the 11 rotated factors and the relationship between classification by the first principal factor and by the rotated factors.

The second principal factor was also used to classify cases according to the 60% criterion, and 49 cases were found to meet this criterion. Figure 4 indicates that classification by the first principal factor was roughly orthogonal to that by the second principal factor.

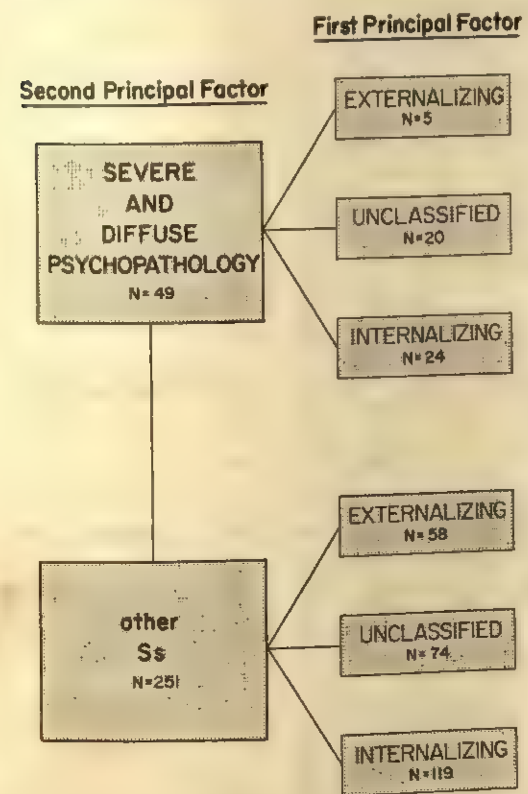

FIG. 4. Classification of female *Ss* by first and second principal factors.

Relationship of Biographical Variables to Factorial Classifications[5]

The Internalizing versus Externalizing classification would appear to offer the most meaningful comparisons on the biographical variables because of the mutually exclusive categories it provides and the large number of *Ss* which fall into each category. For virtually all variables where Internalizers differed significantly from Externalizers, the distribution of Unclassified *Ss* fell midway between the other two.

Since rough measures of "social competence" have been found to be related to the symptom patterns of adult psychiatric patients (Phillips & Zigler, 1961; Zigler & Phillips, 1960) and to personality variables in psychiatric and nonpsychiatric patients (Achenbach & Zigler, 1963), the present study sought to analyse the relationship between various social competence indicators and the juvenile symptom patterns found. The studies just cited usually employed the variables of occupation, age, education, marital status, employment history, and intelligence as social competence indicators. However, there are clearly at least two general variables indicative of sociocultural attainment represented by this type of index. The occupation and education scores represent the traditional concept of social class, while marital status and employment history may be rough indicators of personality adjustment. Age and intelligence probably correlate with both of these general indicators of different types of sociocultural attainment. While the concept of social competence in developmental theory assumes that all six variables correlate in varying degrees with an underlying pattern of adaptive adequacy, it was considered useful for the study of child symptoms to separate social class from adaptive adequacy. Because the social class of the child would have been the product of his parents' behavior and not his own, it could not be regarded as a measure of his adaptive adequacy.

[5] All p values are for two-tailed tests; chi-square tests are presented in Table 5 and all 2 × 2 chi-square values are corrected for continuity; t tests are presented in Table 6; *N*s vary because not all records provided data on all variables.

TABLE 5

CHI-SQUARE TESTS

Variable	Males						Females					
	Internalizing	Externalizing	Totals	df	X^2	P	Internalizing	Externalizing	Totals	df	X^2	P
1. School performance												
Below average	17	62	79	2	21.806	<.001	20	28	48	2	19.657	<.001
Average	15	25	40				47	16	63			
Above average	13	3	16				33	7	40			
Total	45	90	135				100	51	151			
2. Child's previous problems												
(a) Number:												
None	54	45	99	1	33.442	<.001	108	29	137	1	17.590	<.001
1 or more	12	78	90				24	28	52			
Total	66	123	189				132	57	189			
(b) Expelled from school												
No	65	93	158	1	14.766	<.001	131	49	180		expect.	
Yes	1	30	31				1	8	9		*f* too	
Total	66	123	189				132	57	189		small	
(c) Police												
No	64	88	152	1	16.058	<.001	130	41	171	1	29.569	<.001
Yes	2	35	37				2	16	18			
Total	66	123	189				132	57	189			
(d) Psychiatric												
No	64	112	176	1	1.512	<.30	127	51	178		expect.	
Yes	2	11	13				5	6	11		*f* too	
Total	66	123	189				132	57	189		small	
(e) School failure												
No	55	83	138	1	4.704	<.05	113	50	163	1	<1	n.s.
Yes	11	40	51				19	7	26			
Total	66	123	189				132	57	189			
3. Number of parental social problems												
(a) Father												
0	35	37	72	2	8.674	<.02	51	14	65	2	5.360	<.07
1	15	38	53				43	20	63			
More than 1	7	24	31				21	16	37			
Total	57	99	156				115	50	165			
(b) Mother												
0	46	42	88	2	19.797	<.001	73	18	91	2	14.412	<.001
1	11	49	60				31	27	58			
More than 1	5	17	22				8	8	16			
Total	62	108	170				112	53	165			
(c) Both parents combined[a]												
0	37	25	62	3	27.844	<.001	46	12	58	3	13.714	<.01
1	9	16	25				26	6	32			
2	10	52	62				32	20	52			
More than 2	9	27	36				20	19	39			
Total	65	120	185				124	57	181			
4. Parental problem categories												
(a) Father												
Alcoholism												
No	45	71	116	1	<1	n.s.	86	32	118	1	1.495	<.30
Yes	12	28	40				29	18	47			
Total	57	99	156				115	50	165			
Divorce												
No	51	85	136	1	<1	n.s.	105	35	140	1	10.702	<.01
Yes	6	14	20				10	15	25			
Total	57	99	156				115	50	165			

TABLE 5—Continued

Variable	Males						Females					
	Internalizing	Externalizing	Totals	df	X^2	P	Internalizing	Externalizing	Totals	df	X^2	P
Psychiatric history												
No	52	80	132	1	2.270	<.20	90	45	135	1	2.487	<.20
Yes	5	19	24				25	5	30			
Total	57	99	156				115	50	165			
Unemployment												
No	52	76	128	1	4.201	<.05	96	41	137	1	<1	n.s.
Yes	5	23	28				19	9	28			
Total	57	99	156				115	50	165			
(b) Mother: divorce												
No	55	77	132	1	5.915	<.02	97	37	134	1	5.596	<.02
Yes	7	31	38				15	16	31			
Total	62	108	170				112	53	165			
Illegitimate child												
No	59	94	152	1	2.109	<.20	106	49	155		expect.	
Yes	3	14	17				6	4	10		f too	
Total	62	108	170				112	53	165		small	
Psychiatric history												
No	53	82	135	1	1.655	<.20	86	40	126	1	<1	n.s.
Yes	9	26	35				26	13	39			
Total	62	108	170				112	53	165			
5. Lives with												
Both natural parents	52	76	128	1	4.999	<.05	97	33	130	1	3.845	<.05
Other	16	52	68				46	30	76			
Total	68	128	196				143	63	206			
6. Parents' attitude toward child's problem												
(a) Father												
Concerned	38	49	87	1	6.812	<.01	65	22	87	1	9.885	<.01
Resentful or indifferent	12	45	57				21	25	46			
Total	50	94	144				86	47	133			
(b) Mother												
Concerned	61	87	148	1	9.202	<.01	117	40	157	1	14.153	<.001
Resentful or indifferent	4	30	34				13	20	33			
Total	65	117	182				130	60	190			
7. Number of siblings												
(a) 0	4	5	9	6	11.359	<.10	7	3	10	6	6.923	<.50
1	17	15	32				33	14	47			
2	13	27	40				28	11	39			
3	16	35	51				25	18	43			
4	11	15	26				15	7	22			
5	4	11	15				10	6	16			
More than 5	3	20	23				25	4	29			
Total	68	128	196				143	63	206			
(b) 0 or 1	21	20	41	1	5.361	<.05	40	17	57	1	<1	n.s.
More than 1	47	108	155				103	45	148			
Total	68	128	196				143	62	205			
8. Birth order												
First	25	47	72	3	<1	n.s.	47	23	70	3	2.905	<.50
Second	23	40	63				43	17	60			
Middle	14	31	45				39	12	51			
Last (except second)	6	10	16				14	10	24			
Total	68	128	196				143	62	205			
9. Hometown size												
0–1000	5	17	22	3	2.728	<.50	20	4	24	3	8.704	<.05
1001–5000	7	12	19				30	9	39			
5001–25,000	12	29	41				29	8	37			
Above 25,000	44	70	114				64	42	106			
Total	68	128	196				143	63	206			

TABLE 6

t TESTS

Variable	Males							Females						
	Internalizing		Externalizing					Internalizing		Externalizing				
	Mean	*SD*	Mean	*SD*	*df*	*t*	*p*	Mean	*SD*	Mean	*SD*	*df*	*t*	*p*
1. WISC and Binet IQ	104.08	14.04	99.42	12.17	135	1.939	<.10	102.07	15.62	100.16	15.10	125	<1	n.s.
2. Verbal WISC	102.93	15.35	96.61	12.21	89	2.119	<.05	99.62	13.52	94.90	12.54	93	1.381	<.20
3. Performance WISC	101.34	9.76	102.84	13.75	89	<1	n.s.	104.01	14.75	96.79	14.05	87	1.910	<.10
4. Performance minus Verbal for Internalizing	$\overline{D}$ = .97	14.14			28	<1	n.s.	$\overline{D}$ = 4.84	12.21			69	3.317	<.01
5. Performance minus Verbal for Externalizing			$\overline{D}$ = 6.05	11.49	59	4.099	<.001			$\overline{D}$ = 1.90	14.87	18	<1	n.s.
6. Parents' age when *S* was born:														
(a) Father	33.95	7.19	30.68	7.63	179	2.812	<.01	32.77	8.48	30.58	9.57	185	1.549	<.20
(b) Mother	30.30	5.66	24.25	5.84	187	6.885	<.001	27.73	6.04	26.47	5.78	192	1.356	<.20
7. Social class Index I	3.84	1.37	3.48	1.44	193	1.685	<.10	3.32	1.35	3.06	1.48	196	1.218	>.20
8. Social class Index II	3.76	1.47	3.41	1.49	193	1.587	<.20	3.27	1.54	3.00	1.41	196	1.225	>.20
9. Age	11.38	2.71	10.05	3.16	194	2.943	<.01	11.30	2.93	11.38	3.42	204	<1	n.s.

For both the children and their parents, rough measures of adaptive adequacy were recorded. For the children, school performance was one obvious indicator of adaptive adequacy. If the child was regularly attending school, reports of his school performance were recorded in the categories of "below average," "average," and "above average." A chi-square comparing these three categories revealed that Internalizers of both sexes were performing significantly better in school ($\chi^2 = 21.806$, $df = 2$, $p < .001$ for males; $\chi^2 = 19.657$, $df = 2$, $p < .001$ for females). Comparisons of IQs were also made by using full scale Wechsler Intelligence Scale for Children (WISC) scores when available or Stanford-Binet scores in lieu of the WISC. There was no significant difference in IQ between Internalizers and Externalizers of either sex, although the mean IQs for Internalizers of both sexes were higher and the male difference nearly reached significance ($t = 1.939$, $df = 135$, $p < .10$ for males; $t < 1$, $df = 125$, $p =$ n.s. for females). A further analysis of IQ scores was made by comparing the verbal and performance scores on the WISC. For males, Internalizers had significantly higher verbal IQ scores than did Externalizers ($t = 2.119$, $df = 89$, $p < .05$), but there were no significant differences between the verbal scores for female Internalizers and Externalizers, nor between the performance scores of the two groups for either sex. A direct test of the difference between *S*'s verbal and performance scores showed male Externalizers to have significantly higher performance than verbal scores ($t = 4.099$, $df = 59$, $p < .001$), while female Internalizers also had significantly higher performance than verbal scores ($t = 3.317$, $df = 69$, $p < .01$). There were no significant differences between performance and verbal scores for male Internalizers or female Externalizers.

The number of different categories of social problems which *S* had been reported to manifest constituted an additional index of adaptive adequacy. The categories of problems recorded were: (*a*) trouble with the police and being brought to court, (*b*) previous psychiatric referral, (*c*) being expelled from school, and (*d*) failing a grade in school. No measure of the degree of difficulty in any of the categories was attempted, but it was assumed that the number of categories in which entries occurred would contribute a rough measure of the child's previous social adequacy. As the Externalizing pole that appeared in the factor analyses included several behaviors which could have resulted in entries in some of the problem categories, the basis for this index was not entirely independent of the Internalizing-Externalizing classification. Its problem categories may be best regarded, perhaps, as general consequences of the type of adaptive pattern of which the Externalizing symptoms are specific elements. Internalizers of both sexes had far fewer of these problems than did Externalizers ($\chi^2 = 33.442$, $df = 1$, $p < .001$ for males; $\chi^2 = 17.590$, $df = 1$, $p < .001$ for females). For all categories and each sex, the Externalizers had proportionally more entries than did the Internalizers, although not all the differences were significant. For the males, the differences were significant in three of the categories (police, $\chi^2 = 16.058$, $df = 1$, $p < .001$; school failure, $\chi^2 = 4.704$, $df = 1$, $p < .05$; expelled from school, $\chi^2 = 14.766$, $df = 1$, $p < .001$). In the fourth category, previous psychiatric referral, the expected frequency in one cell reached only 4.54, so the χ^2 of 1.512 is of questionable validity, but is far from significant. For the females, frequencies in the psychiatric and expelled categories were too small for analysis. Externalizers had significantly more trouble with the police ($\chi^2 = 29.569$, $df = 1$, $p < .001$), but there was not a significant difference in the school failure category ($\chi^2 < 1$).

The adaptive adequacy index for the parents was composed of the following problem categories: (*a*) divorce, (*b*) psychiatric history, (*c*) criminal record, (*d*) frequent excessive use of alcohol, (*e*) unemployment, (*f*) having an illegitimate child, (*g*) desertion of family, and (*h*) being charged with neglect of children. Chi-square tests revealed that fathers of Externalizers of both sexes manifested more of these problems, but the difference for females did not quite reach statistical significance ($\chi^2 = 8.674$, $df = 2$, $p < .02$ for the males; $\chi^2 = 5.360$, $df = 2$, $p < .07$ for the females). Mothers

of Externalizers of both sexes also manifested more problems, the difference being highly significant in each case ($\chi^2 = 19.797$, $df = 2$, $p < .001$ for males; $\chi^2 = 14.412$, $df = 2$, $p < .001$ for females). The sum of problems for both parents, or twice the score of one parent if there were no data for the other parent, showed the parents of Externalizers of both sexes to have significantly more problems ($\chi^2 = 27.844$, $df = 3$, $p < .001$ for males; $\chi^2 = 13.714$, $df = 3$, $p < .01$ for females).

A breakdown of the specific categories of problems showed that fathers of male Externalizers tended to have proportionately more entries in all categories except that of illegitimate children. However, these differences were significant only for the category of unemployment ($\chi^2 = 4.201$, $df = 1$, $p < .05$) and nonsignificant for divorce, alcohol, and psychiatric history. In the categories of criminal history, illegitimate children, desertion, and neglect, the frequencies were too small for valid chi-square analysis. The fathers of female Externalizers tended to have proportionally more entries in the categories of divorce, criminal history, alcohol, desertion, and neglect. The difference was significant only for the category of divorce ($\chi^2 = 10.702$, $df = 1$, $p < .01$) and nonsignificant for the category of alcoholism. Frequencies were too small for chi-square analysis in the categories of criminal history, illegitimate children, desertion, and neglect.

Mothers of Externalizers of both sexes tended to have proportionally more entries in the six categories where mothers had any entries at all (there were none in the criminal or unemployment categories). These differences were significant for the divorce category ($\chi^2 = 5.915$, $df = 1$, $p < .02$ for males; $\chi^2 = 5.596$, $df = 1$, $p < .02$ for females), but nonsignificant for psychiatric history for the mothers of both sexes and for illegitimate children for the mothers of males. All other categories contained frequencies too small for chi-square analysis.

A comparison of the persons with whom Internalizers and Externalizers were residing showed that Internalizers of both sexes were more frequently living with both natural parents than were Externalizers ($\chi^2 = 4.999$, $df = 1$, $p < .05$ for males; $\chi^2 = 3.845$, $df = 1$, $p < .05$ for females). Where sufficient data were available, the attitude of each parent or parent surrogate toward having the child's problem treated was rated in the categories of: (*a*) resentful; (*b*) indifferent; and (*c*) concerned. Chi-square comparisons showed that both parents of Internalizers of both sexes were more often rated "concerned" than were parents of Externalizers ($\chi^2 = 6.812$ for fathers, and $\chi^2 = 9.202$ for mothers, both $df = 1$, $p < .01$ for males; $\chi^2 = 9.885$, $df = 1$, $p < .01$ for fathers, $\chi^2 = 14.153$, $df = 1$, $p < .001$ for mothers of females).

A *t*-test comparison of the age of each parent when the child was born showed that both parents of male Internalizers were significantly older than those of male Externalizers ($t = 2.812$, $df = 179$, $p < .01$ for fathers; $t = 6.885$, $df = 187$, $p < .001$ for mothers, but that there was only a nonsignificant trend in the same direction for females ($t = 1.549$, $df = 185$, $p < .20$ for fathers; $t = 1.356$, $df = 192$, $p < .20$ for mothers).

In order to examine the relationship between the social class component of social competence and the symptom patterns, a measure of parental social class obtainable from the case histories was necessary. In the Zigler-Phillips studies, the six social competence variables were scaled on equivalent scales. Those variables for which data were present in a given case history were averaged to yield the social competence score. For present purposes, where only a social class score rather than an overall social competence score was sought, a somewhat similar approach was taken. The six-step scale for education and a modification of the six-step scale for occupation employed by Zigler and Phillips (1960), with occupation being classified by the *Dictionary of Occupational Titles* (United States Government, 1949), were used to calculate the social class score for the head of the household. The education and occupation scores were averaged to yield the social class score. Where only one variable was reported, it provided the social class score.

Cases where neither occupation nor education were given, but where the family depended upon public welfare assistance were automatically assigned to the lowest category. Table B1 presents the scales and the distribution of social class scores obtained by this method (Social Class Index I). There was a nonsignificant tendency for Internalizers of both sexes to have higher social class scores than did Externalizers ($t = 1.685$, $df = 193$, $p < .10$ for males; $t = 1.218$, $df = 196$, $p < .20$ for females).

The scaling approach just described assumed that the averaging of available data gave roughly consistent estimates of social class standing, and it resulted in a relatively continuous distribution of scores. Another approach to social class scaling is by attempting to assign individuals to a few discrete strata. Hollingshead and Redlich (1958), for example, employed the variables of occupation, education, and neighborhood to assign individuals to one of five classes which were thought to exist in the community. To see if this approach would yield results different from those reported above, cases were classified using a modification of Hollingshead's "Two-Factor Index of Social Position" (1957). Normally, the two scores from Hollingshead's seven-step occupation scale and seven-step education scale are averaged, with occupation being given a weight of seven and the education score a weight of four. To obtain a small number of discrete social classes, only an occupation score (found by Hollingshead to be the best predictor of social class) was employed here; when it was not available, the education score was used instead; if both were unknown and the family was on welfare, the case was assigned to the lowest class. Except for the placement of service workers and the combining of the top two occupational and educational classes into one, the six-step scales presented with Table B1 resemble the Hollingshead scales rather closely. Therefore, in this second social class scaling procedure (Social Class Index II), the occupational and educational categories of Table B1 were followed, but the specific occupational roles of service workers were scored according to the Hollingshead Index. For example, a practical nurse, classified as a personal service worker by the *Dictionary of Occupational Titles*, would have been assigned to Category 4 by the scale in Table B1, but was listed by Hollingshead under the same class heading as "Machine Operators and Semi-skilled Employees." Several other specific occupational roles listed as service workers, for example, policemen, firemen, nightwatchmen, janitors, and waiters, were also classified differently under the Hollingshead Index. As with Social Class Index I, there were nonsignificant tendencies for Internalizers of both sexes to have the higher social class scores ($t = 1.587$, $df = 193$, $p < .20$ for males; $t = 1.225$, $df = 196$, $p > .20$ for females).

The distributions of diagnoses are presented in Table B3. Examination of the remaining biographical variables showed the following results: Internalizing males were significantly older than Externalizing males ($t = 2.943$, $df = 194$, $p < .01$), but there was not a significant difference in age between Internalizing and Externalizing females ($t < 1$, $df = 204$, $p =$ n.s.; age distributions are presented in Table B2); there was a tendency, approaching significance, for Externalizing males to have more siblings than did Internalizing males, but the difference for the females was in the opposite direction and insignificant ($\chi^2 = 11.359$, $df = 6$, $p < .10$ for males; $\chi^2 = 6.923$, $df = 6$, $p < .50$ for females); 2×2 chi-squares comparing the number of cases who had zero or one sibling with those having more siblings showed that Externalizing males more frequently had more than one sibling ($\chi^2 = 5.361$, $df = 1$, $p < .05$; $\chi^2 < 1$, $df = 1$, $p =$ n.s. for females); there were no significant differences in birth order for either sex ($\chi^2 < 1$, $df = 3$, $p =$ n.s. for males; $\chi^2 = 2.905$, $df = 3$, $p < .50$ for females); there was no significant difference in hometown size for the two male groups ($\chi^2 = 2.728$, $df = 3$, $p < .50$), but female Externalizers came from larger towns than did female Internalizers ($\chi^2 = 8.704$, $df = 3$, $p < .05$); there appeared to be no consistent religious differences in the groups and no grounds for combining low frequency categories to make a chi-square test possible.

No independent analysis of the background data for *S*s classified by the second principal factor was attempted because of the general weakness of that factor as compared to the first principal factor. Inspection of biographical data for *S*s classified by the rotated factors[6] revealed that the female Factor 11, Enuresis and Other Immaturities, represented only very young children. Five of the seven *S*s were 4 years old and two were 5, yielding a mean age of 4.29. For both sexes, the Hyperreactive Behavior factor also classified predominantly younger *S*s, the mean ages being 7.33 for the females and 7.80 for the males. The *S*s classified by Factor 3, Obsessions, Compulsions, and Phobias, had the highest mean IQ for each sex (111.89 for males, 114.67 for females), while *S*s classified by the Aggressive Behavior factor had the lowest mean IQ for the females (86.50) and the second lowest for the males (95.71). The same relationships held for social class, with the Obsessive *S*s being the highest ($M = 4.46$ for males; $M = 3.96$ for females), and the Aggressive *S*s being the lowest for each sex ($M = 3.33$ for males; $M = 2.50$ for females). Finally, the parents of the Aggressive *S*s had the highest mean numbers of social problems in each sex group (3.00 for males; 2.50 for females).

Discussion

With regard to the first purpose of the study, the data indicate that child psychiatric symptoms do indeed form both general clusters like those found for adults by Phillips and Rabinovitch and Guertin, and for children by Hewitt and Jenkins, and more specific clusters like those implied by traditional diagnostic categories and found for adults by Wittenborn. The types of symptoms falling at the Internalizing pole of the first principal factor for both sexes are certainly consistent with those of Hewitt and Jenkins' Overinhibited Child cluster and, except for the obvious developmental differences, the symptoms of Phillips and Rabinovitch's "self-deprivation and turning against the self" cluster and Guertin's Guilt-Conflict cluster. Likewise, the symptoms at the Externalizing pole of the first principal factor for both sexes are consistent with the Hewitt-Jenkins Socialized Delinquent and Unsocialized Aggressive clusters, the Phillips-Rabinovitch "self-indulgence and turning against others" cluster, and the Guertin Excitement-Hostility cluster.

Many of the rotated factors which were considered reliable resemble traditional categories and the factors found by Wittenborn. The Somatic Complaints, Obsessions, Compulsions, and Phobias, and Schizoid Thinking and Behavior factors found for both sexes are reminiscent of Wittenborn's "conversion hysteria," "phobic compulsive," and "schizophrenic excitement" syndromes. The Anxiety Symptoms and Depressive Symptoms factors found here for the girls resemble his "acute anxiety" and "depressed state" syndromes, respectively. Since his sample excluded sociopaths, syndromes corresponding to the present Delinquent Behavior and Aggressive Behavior factors would not have been expected to occur, but these may well be the child counterparts of the diagnostic categories of Dyssocial Reaction and Antisocial Reaction, respectively. The Enuresis and Other Immaturities factor found for the girls is evidently a syndrome peculiar to early childhood (occurring only in 4- and 5-year olds in the present sample), and would not be expected in adult populations. The Hyperreactive Behavior factor also classified only young children in the present sample and may belong to an early developmental stage or may be the early sign of what is later recognized as organic dysfunction, causing such individuals to be then excluded from functional categories.

The Obesity factor classified only girls aged 10–14 and might therefore be regarded as a developmental phenomenon associated

[6] Two one-page tables presenting mean age, IQ, parental problems, and social class scores of *S*s classified by the male and female rotated factors have been deposited with the American Documentation Institute. Order Document No. 8848 from ADI Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance $1.25 for microfilm or $1.25 for photocopies and make checks payable to: Chief, Photoduplication Service, Library of Congress.

with puberty and unlikely to occur independent of other syndromes in an adult sample. Likewise, the vast majority of *S*s classified by the Neurotic and Delinquent Behavior factor were between the ages of 12 and 15, suggesting that this too is a phenomenon belonging to a specific developmental period and not to be expected in patients from other age groups. In passing, it is to be noted that in the male data factors similar to the female Depressive Symptoms and Obesity factors occurred in several rotations, but that they failed to meet the criterion of reliability employed. While the factors concluded to be reliable can probably be regarded with a good deal of confidence, they cannot necessarily be regarded as exhausting the child-symptom domain. The evidence indicates however, that general symptom clusterings, here labeled Internalizing versus Externalizing, and specific syndromes like several of the traditional syndromes of adult diagnosis exist in the child domain. In addition, there is evidence for several syndromes which are apparently peculiar to specific childhood age periods and which are not recognized in adult diagnosis.

The discovery of numerous reliable factors beside the major Internalizing-Externalizing dichotomy means that the second purpose of the study, that of obtaining a more differentiated operational classification schema for research purposes, has been in part fulfilled. The factorial results indicate that there are indeed several discrete and reliable clusterings of symptoms. The best means for classifying patients using these clusterings will depend upon the exact research purposes for which such a classification schema is to be used and the heuristic value which it is found to possess. It would appear that the major Internalizing-Externalizing dichotomy can be readily used for classifying both case histories and live patients in the same way as it was used here. If 60% or more of a case's symptoms came from the Internalizing cluster reported in Table 1 for males or Table 3 for females, the case would be assigned to the Internalizing category, and likewise for the Externalizing category. If only a few cases were available, a more liberal criterion, such as a simple majority of symptoms, could be employed to classify them. If subtle differences were being investigated, a more rigorous criterion, such as 75% or 100%, could be employed.

The classification of cases by the rotated factors in Tables 2 and 4 could also be done by the 60% criterion used here. However, since the *N*s classified here by some of the rotated factors were quite small, a more liberal criterion might generally be needed. An obvious alternative to classification by means of the percentage of symptoms matching a factor would be to assign *S*s to a category if they manifested all or most of the most heavily loaded symptoms on the factor. Another approach would be to classify *S*s by the Internalizing and Externalizing clusters, and also by those rotated factors which were not clearly subsumed by the Internalizing and Externalizing clusters, for example, the Schizoid and Hyperreactive factors.

Regarding the third purpose of the study, the classification of individual cases by their degree of resemblance to both the rotated and unrotated factors revealed similar relationships among the groupings for both sexes (see Figures 1 and 3). All *S*s classified by the Aggressive Behavior and Delinquent Behavior factors had already been classified by the Externalizing end of the first principal factor. Virtually all *S*s classified by the Somatic Complaints and Obsessions, Compulsions, and Phobias factors had already been classified by the Internalizing end of the first principal factor. Individuals classified by the Hyperreactive Behavior factor came from both the Externalizing and Unclassified groups of the first principal factor, while individuals classified by the Schizoid factor tended to come from the Internalizing and Unclassified groups, with a few males coming from the Externalizing group. For both sexes then, the Aggressive and Delinquent factors may be regarded as representing subtypes within the general Externalizing category, while the Somatic and Obsessive factors represent subtypes within the general Internalizing category. The Hyperreactive and Schizoid factors

may be regarded as not being clearly subsumed by the categories of the first principal factor, although the Hyperreactive factor clearly excludes Internalizers and the Schizoid factor tends to exclude Externalizers. Of the factors peculiar to the females, the Anxiety factor appears clearly subsumed by the Internalizing category, while the Depressive and Obesity factors are not clearly subsumed by it but exclude Externalizers. The Neurotic and Delinquent factor, on the other hand, is not clearly subsumed by the Externalizing category, but tends not to include *S*s classified as Internalizers. Thus the relationship between the classification of *S*s by the general categories and by the discrete syndromes is in part an hierarchical one, with several of the discrete syndromes representing clear subcategories of the general clusters, while several other of the discrete syndromes appear related to two of the general groupings, but are not neatly subsumed by a single one. The classification of *S*s by the second principal factor for each sex was orthogonal to classification by the first principal factor (see Figures 2 and 4) and was found to bear no consistent relationship to any of the rotated factors.

For both sexes, the Internalizing-Externalizing dichotomy significantly discriminated cases on many of the biographical variables. The findings that Internalizers were more frequently living with both natural parents and that their parents had fewer overt social problems than the parents of the Externalizers corroborates the findings of Bennett with regard to neurotic and delinquent children and Hewitt and Jenkins with regard to Overinhibited versus Socialized Delinquent and Unsocialized Aggressive children. The greater frequency with which parents of Internalizers were rated "concerned," as contrasted with "resentful" or "indifferent," further suggests that the Internalizers' parents took more responsibility for their children than did Externalizers' parents. The findings that Internalizers of both sexes had significantly fewer previous social problems and significantly better school performance suggests that, like their parents, the Internalizers had been socially more adequate prior to their psychiatric referral than were the Externalizers.

An interesting secondary finding was that Externalizing males and Internalizing females had significantly higher performance than verbal IQ scores on the WISC. It has previously been found (e.g., Glueck & Glueck, 1950, 1959) that delinquent boys have significantly higher performance than verbal IQ scores, but evidently this relationship has not been investigated for girls. If the finding that Internalizing girls have higher performance than verbal IQs is a reliable one, previous speculation about lack of symbolic ability in delinquent boys being related to their delinquent behavior may have to be reconsidered.

The heuristic goal of the present study rests upon the interpretations of the empirical relationships discovered. To what extent can the observed relationships between the two levels of factorial classification and the relationships between the factors and the biographical variables lead to conceptual order and the eventual generation of testable hypotheses? First, the grouping together of the *S*s classified by the Somatic, Obsessive, and Anxiety Factors under the Internalizing category implies that these *S*s have something in common which is reflected in the functional unity of the Internalizing cluster. On the other hand, the Aggressive and Delinquent *S*s have something in common which is reflected in the functional unity of the Externalizing cluster. Second, the Hyperreactive, Schizoid, and Depressive factors appear not to be directly related to the functional unities embodied in the first principal factor. Third, the second principal factor appears to be irrelevant to the groupings defined by the first principal factor and the rotated factors.

It is clear that the Externalizing symptoms for both sexes represent behavior which is antisocial and which most people learn through negative sanctions not to perform. For the behavior theorist, the obvious interpretation of the common feature in the *S*s classified as Externalizers is that counterconditioning of antisocial behavior has not been successful. Bandura and Walters

(1959) concluded that one of the major differences between normal and aggressive adolescents was that antisocial behavior in normal adolescents was prevented by guilt feelings whereas aggressive adolescents showed an absence of guilt and were deterred from misbehaving only when fear-arousing consequences were evident in the immediate situation. Bandura and Walters interpreted this to mean that successful socialization requires the substitution of less aggressive new responses for overt aggression, rather than the displacement of the original responses onto new objects (p. 139). The findings of high frequencies of overt social problems and lack of parental concern in the family backgrounds of the Externalizers indicates that the social learning regimes they experienced probably did not provide the combination of reward contingencies and good role models which are necessary both to deter antisocial responses and to promote socialized responses. Following this line of reasoning, one could conclude that the deviant nature of the behavioral reactions manifested by Externalizers inheres in the absence of the proper learning of socialized behavior. The deviant behavior manifested by the Internalizers, on the other hand, presupposes the acquisition, through a socialization process, of behavioral reactions which are not antisocial.

If the foregoing interpretation is accepted, it can be said that the Aggressive and Delinquent syndromes simply represent subvarieties of individuals whose social learning regimes have not successfully eliminated antisocial behavior. The Somatic, Obsessive, and Anxiety syndromes represent individuals whose social learning regimes have promoted adaptive patterns which are more socialized. What then of the Unclassified *S*s and those represented by the Schizoid, Hyperreactive, Neurotic and Delinquent, and Depressive syndromes? First, it would appear that some of the Unclassified *S*s have mixtures of Externalizing and Internalizing symptoms, whereas others tend to have symptoms which were loaded on neither the Externalizing nor Internalizing poles. The *S*s classified by the Enuresis and Other Immaturities factor may be of this latter group, since nine of the 13 symptoms on this factor had loadings smaller than $\pm.200$ on the principal factor and five of the seven *S*s classified by it came from the Unclassified group, with one each from the Externalizing and Internalizing groups. Second, it would appear that some individuals who belong to the Schizoid, Hyperreactive, Neurotic and Delinquent, and Depressive groups clearly belong to either the Externalizing or Internalizing groups, whereas others are from the Unclassified group. It might therefore be concluded that while some of these individuals have things in common with Internalizers or Externalizers, the presence or absence of antisocial behavior is not a defining characteristic of their syndrome. For example, some of the Schizoid *S*s may be Externalizers because the socialization process has not successfully eliminated antisocial behavior, while others are Internalizers because antisocial behavior has been successfully eliminated, but what they have in common is orthogonal to the Internalizing-Externalizing (or "socialization") dimension. Such a conclusion would be compatible with both organic and psychodynamic theories of schizophrenia which postulate a fundamental defect in the capacity for integrated behavior—the particular behaviors manifested might be influenced by social learning, but the underlying functional unity which determines the schizoid patterning is independent of socialization.

If, for either psychological or organic reasons, the young children manifesting the Hyperreactive syndrome were unable to inhibit impulse or to integrate their behavior, a variety of behavior might be expected to result which, while not readily suppressible by socialization, was not necessarily always antisocial. The Depressive syndrome classifies individuals from both the Internalizing and Unclassified groups. This may imply that, while not all *S*s in this group are Internalizers, the condition is unlikely, perhaps because of general retardation in functioning, to be accompanied by much antisocial behavior.

According to the above reasoning, the social adjustment component of the social

competence concept would refer to the presence of socialized adaptive patterns due to learning. Insofar as symptoms are purely "behavioral reactions," those manifested by children and adults who have been well socialized should differ from those who have not been socialized in that the former will not include antisocial behavior. On the other hand, the social class component of adult social competence may measure the resultant of both a socialized adaptive pattern and the intelligence and social opportunities which are needed for various occupational levels. If socialized adaptive patterns and intelligence are both necessary to the maintenance of high social class standing by adults in a society which allows upward and downward mobility, the correlation between these two variables and social class should increase with age, as an individual's social class standing comes to depend more upon his own behavior and less directly on that of his parents. In an adult psychiatric population, such as that employed in the Phillips-Zigler studies, the intercorrelation of social class (occupation-education), social adjustment (marital status–employment history), age, and intelligence should be high and may well reflect a unitary underlying pattern of adult adaptive adequacy. However, in the present data, the premorbid adequacy of children and their parents were significantly related to the Internalizing-Externalizing dimension in symptoms, but intelligence of the child and social class of the parents showed only nonsignificant relationships to the Internalizing-Externalizing dimension. This suggests that the Internalizing symptom pattern and the prevention of antisocial behavior depend more upon parents who themselves do not display antisocial behavior than upon the intelligence of the child or the social class correlates of his environment.

A final qualification to the above interpretation of the factorial classification must be added. The factors are based only upon symptoms which are defined by relatively peripheral behavior, rather than by experiental or central variables. The factors are thus no more than collections of observed behaviors. Another approach to the statistical classification of psychiatric disorders has been by the intercorrelation of personality test scores. Eysenck (cf. Eysenck, 1961, pp. 1–31) has pursued this approach most extensively and has repeatedly found two dimensions which he has labeled Introversion-Extraversion and Neuroticism. Psychiatric patients diagnosed hysterical or psychopathic have been found to have factor scores high on Neuroticism and high on Extraversion; patients diagnosed anxious or obsessional have been high on Neuroticism and high on Introversion. Insofar as the present Somatic, Delinquent, Anxiety, and Obsessive *S*s correspond to Eysenck's diagnostic groups, there is an evident contrast in that Somatic, Anxiety, and Obsessive *S*s were found here to group together at what has been called the Internalizing pole of the first principal factor whereas Delinquent *S*s fell at the Externalizing pole. Since the present data consisted of child symptoms and Eysenck's data were adult test scores, the findings are not directly contradictory. In fact, further consideration suggests that they may be complementary. If the symptoms of these four groups are regarded as learned behavior, and the third dimension of successful socialization versus unsuccessful socialization is added to Eysenck's two dimensions, it can be seen that, while psychopaths and hysterics would be similar in being extroverted, hysterics would be similar to anxiety and obsessional *S*s in not manifesting antisocial behavioral reactions. There might thus be three independent dimensions along which psychiatric disorders of these four types could be classified: Eysenck's two dimensions assume genetic predispositions ("inherited autonomic over-reactivity" for Neuroticism, and "strong conditionability" for Introversion, p. 21), while the present dimension involves a fundamental distinction in patterns of socialization.

Eysenck has also produced evidence for a psychoticism dimension in test scores. If such a dimension is regarded simply as one of disturbance not readily shaped by socialization, it might be relevant to the Schizoid and Hyperreactive factors found here, or to

the second principal factor. Another possible interpretation of the second principal factor is that it represents the intensity of disturbance and classifies very disturbed individuals regardless of the type of syndrome manifested.

SUMMARY AND CONCLUSIONS

Symptoms and biographical data were recorded from the case histories of 300 male and 300 female child psychiatric patients. The symptoms were intercorrelated and factor analyzed, separately for each sex, by the principal factor method, and the factors were rotated to the varimax, quartimax, and oblimin criteria for simple structure. The first principal factor for both sexes was bipolar, with antisocial behavior ("Externalizing") at one end and symptoms of internal problems ("Internalizing") at the other end. Several factors occurred repeatedly in the different rotations and were therefore considered reliable. For both sexes, factors given the labels of Somatic Complaints, Obsessions, Compulsions, and Phobias, Delinquent Behavior, Aggressive Behavior, Hyperreactive Behavior, and Schizoid Thinking and Behavior were found. For the boys alone, a factor labeled Sexual Problems, and, for the girls alone, factors labeled Depressive Symptoms, Anxiety Symptoms, Neurotic and Delinquent Behavior, Enuresis and Other Immaturities, and Obesity were found.

Cases were classified according to their resemblance to the poles of the principal factor and to the rotated factors. If 60% or more of an *S*'s symptoms came from a given factor, or pole of the principal factor, he was placed in the category represented by that factor or pole. This revealed that all *S*'s classified by the Aggressive and Delinquent rotated factors of both sexes were also classified by the Externalizing pole of the principal factor. Almost all *S*s classified by the Somatic and Obsessive rotated factors of both sexes and the Anxiety factor of the girls were also classified by the Internalizing pole of the first principal factor. Classification by the other rotated factors bore no consistent relation to one or the other pole of the principal factor.

Because the parents of Externalizers were found to have significantly more overt social problems and to be rated less concerned with their child's difficulty, it was suggested that the Externalizing symptoms reflected a learning regime in the child's home leading to antisocial behavior, while the Internalizing symptoms presupposed the learning of more socialized behavioral reactions. The Aggressive and Delinquent factors were concluded to be subvarieties of the general category of unsocialized behavioral reactions, while the Somatic, Obsessive, and Anxiety factors were concluded to be subvarieties of the general category of socialized behavioral reactions. The other rotated factors were concluded to represent functional unities which were relatively independent of the socialization dimension. Some of these appeared to be peculiar to specific developmental stages.

The overall results showed that the factors obtained can be used directly for the classification of child psychiatric cases for research purposes. The more general and more specific clusterings of symptoms may be employed independently or in an hierarchical ordering. The relationships among the clusterings suggested that different diagnostic models may ultimately be appropriate for different syndromes.

APPENDIX A

SYMPTOMS (OCCURRING WITHIN THREE YEARS OF ADMISSION)[A1]

M, F 1. Apathy, underactive, no initiative, slow, lethargic
2. Assault with weapon
M, F 3. Asthma
M, F 4. Attention demanding
5. Believes he is evil
M, F 6. Bizarre behavior
7. Breathholding
F 8. Breathing difficulty

[A1] M and F indicate symptoms which occurred five or more times in the male and female samples, respectively.

M, F 9. Can't concentrate, distractible, short attention span
10. Claustrophobia
M, F 11. Complains no one loves him, feels rejected
M, F 12. Compulsions
M, F 13. Confused
M, F 14. Constipation
M 15. Cruelty, bullying, meanness
M, F 16. Crying
M, F 17. Daydreaming, excessive fantasy
M, F 18. Depression, unhappiness, sadness
M, F 19. Destructive
20. Diarrhea
M, F 21. Diplopia, blurred vision, tubular vision, microscopia
M, F 22. Disobedient, rebellious, discipline problem
M, F 23. Dizziness
M, F 24. Encopresis, soiling
M, F 25. Enuresis, wetting
F 26. Excessive talking, chattering
F 27. Fainting
M, F 28. Fantastic thinking, delusions, hallucinations
M, F 29. Fearful, anxious
M, F 30. Fears own impulses
M, F 31. Feelings of worthlessness, inadequacy, inferiority
M, F 32. Fighting, assault, aggressive behavior
M 33. Fire-setting
34. Glue-sniffing, addictions
35. Grinding teeth
M, F 36. Headaches
M, F 37. Ideas of reference, feels persecuted, suspicious
M, F 38. Inadequate guilt feelings
F 39. Inappropriately indifferent, e.g., to physical complaints, *la belle indifference*
M, F 40. Insomnia
M, F 41. Loneliness
M 42. Loudness
M, F 43. Lying, cheating
M, F 44. Masochism, self-harm, suicidal, threatens to kill self
M, F 45. Masturbation
M, F 46. Moodiness, rapid change of mood
M, F 47. Nailbiting
M, F 48. Nausea, feels sick
M, F 49. Negativistic, stubborn, sullen, irritable
M, F 50. Nervous, high strung
M, F 51. Nightmares
M, F 52. Obese
M, F 53. Obsessions
F 54. Overeating
M, F 55. Overtired, fatigued, drowsy
M, F 56. Pains, physical complaints
57. Peeping, voyeurism
M, F 58. Phobias, fears
F 59. Picking
M, F 60. Poor motor coordination
M, F 61. Poor school work
M, F 62. Quarrelsome
M, F 63. Refusing to eat, not eating well
M, F 64. Refusing to talk, mute
M, F 65. Restless, hyperactive
66. Ritualistic behavior
M, F 67. Running away
M, F 68. Seclusive
M, F 69. Self-conscious
M, F 70. Sexual delinquency, incest, homosexuality
M 71. Sexual perversions, exposing self
M, F 72. Sexual preoccupation, precociousness
M 73. Showing off
M, F 74. Shy, timid, submissive
75. Silliness
M, F 76. Skin eruptions
M 77. Sleepwalking
78. Smearing feces
M, F 79. Stealing
M, F 80. Stomachache, cramps, abdominal pain
M, F 81. Stuttering, speech problem
M, F 82. Swearing
M, F 83. Temper tantrums
M, F 84. Threatening people
M, F 85. Thumbsucking
M, F 86. Truancy
M, F 87. Tics, trembling, shaking
M 88. Vandalism
M, F 89. Vomiting
M, F 90. Withdrawn
M, F 91. Worrying

APPENDIX B

TABLE B1
SOCIAL CLASS INDEX 1

Social class	Males								Females							
	Internalizing		Unclassified		Externalizing		Totals		Internalizing		Unclassified		Externalizing		Totals	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
1.0	2	3.0	6	5.8	14	10.9	22	7.4	14	10.3	8	8.7	12	19.4	34	11.7
1.5	0	0.0	2	1.9	3	2.3	5	1.7	2	1.5	4	4.3	0	0.0	6	2.1
2.0	9	13.4	11	10.6	13	10.2	33	11.0	16	11.8	19	20.7	11	17.7	46	15.9
2.5	1	1.5	2	1.9	2	1.6	5	1.7	5	3.7	2	2.2	3	4.8	10	3.4
3.0	16	23.9	24	23.1	28	21.9	68	22.7	37	27.2	20	21.7	7	11.8	64	22.1
3.5	2	3.0	4	3.8	6	4.7	12	4.0	4	2.9	4	4.3	3	4.8	11	3.8
4.0	16	23.9	23	22.1	28	21.9	67	22.4	34	25.0	21	22.8	17	27.4	72	24.8
4.5	0	0.0	4	3.8	8	6.3	12	4.0	3	2.2	4	4.3	0	0.0	7	2.4
5.0	10	14.9	16	15.4	13	10.2	39	13.0	10	7.4	4	4.3	5	8.1	19	6.6
5.5	1	1.5	4	3.8	1	0.8	6	2.0	0	0.0	2	2.2	0	0.0	2	0.7
6.0	10	14.9	8	7.7	12	9.4	30	10.0	11	8.1	4	4.3	4	6.5	19	6.6
Totals	67	100.0	104	99.9	128	100.2	299	99.9	136	100.1	92	99.8	62	100.0	290	100.1
No data	1	1.5	0	0.0	0	0.0	1	0.3	7	4.9	2	2.1	1	1.6	10	3.3

Note.—The occupation and education scores of the head of the family were averaged to give the social class score. If only occupation or only education was known, it was used to obtain the social class score.

Occupation scale

1 = unskilled
2 = semiskilled
3 = skilled, domestic service, building service (maintenance)
4 = clerical, minor sales, technical, personal and protective service, owners of farms and little businesses
5 = owners and managers of small businesses, semiprofessionals, major salesmen, administrative personnel of large firms
6 = professional and managerial, owners of medium and large businesses

Education scale

1 = less than eighth grade
2 = completed eighth or ninth grade
3 = completed tenth or eleventh grade
4 = graduated from high school
5 = 1 year or more of college, business college, art school, etc.
6 = completed 4 years of college or more

TABLE B2
AGE

Age	Males								Females							
	Internalizing		Unclassified		Externalizing		Totals		Internalizing		Unclassified		Externalizing		Totals	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
4	1	1.5	3	2.9	7	5.5	11	3.7	1	0.7	4	4.3	4	6.3	9	3.0
5	1	1.5	1	1.0	5	3.9	7	2.3	1	0.7	5	5.3	2	3.2	8	2.7
6	3	4.4	7	6.7	7	5.5	17	5.7	8	5.6	3	3.2	3	4.8	14	4.7
7	1	1.5	7	6.7	9	7.0	17	5.7	9	6.3	7	7.4	2	3.2	18	6.0
8	4	5.9	14	13.5	17	13.3	35	11.7	11	7.7	6	6.4	1	1.6	18	6.0
9	5	7.4	14	13.5	11	8.6	30	10.0	10	7.0	14	14.9	6	9.5	30	10.0
10	8	11.8	10	9.6	15	11.7	33	11.0	17	11.9	11	11.7	2	3.2	30	10.0
11	8	11.8	7	6.7	9	7.0	24	8.0	13	9.1	7	7.4	3	4.8	23	7.7
12	12	17.6	11	10.6	9	7.0	32	10.7	12	8.4	5	5.3	11	17.5	28	9.3
13	9	13.2	16	15.4	16	12.5	41	13.7	18	12.6	7	7.4	5	7.9	30	10.0
14	6	8.8	9	8.7	16	12.5	31	10.3	18	12.6	12	12.8	13	20.6	43	14.3
15	10	14.7	5	4.8	7	5.5	22	7.3	25	17.5	13	13.8	11	17.5	49	16.3
Totals	68	100.1	104	100.1	128	100.0	300	100.1	143	100.1	94	99.9	63	100.1	300	100.1

TABLE B3

DIAGNOSIS

Diagnosis	Males								Females							
	Internalizing		Unclassified		Externalizing		Totals		Internalizing		Unclassified		Externalizing		Total	
	N	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%	*N*	%
1. Adjustment reaction of adolescence	10	14.7	18	17.3	23	18.0	51	17.0	18	12.6	16	17.0	15	23.8	49	16.3
2. Adjustment reaction of childhood	9	13.2	16	15.4	16	12.5	41	13.7	13	9.1	11	11.7	6	9.5	30	10.0
3. Adjustment reaction with habit disturbance	1	1.5	0	0.0	4	3.1	5	1.7	2	1.4	4	4.3	1	1.6	7	2.3
4. Adjustment reaction with conduct disturbance	5	7.4	8	7.7	24	18.8	37	12.3	0	0.0	6	6.4	5	7.9	11	3.7
5. Adjustment reaction with neurotic traits	5	7.4	6	5.8	3	2.3	14	4.7	11	7.7	5	5.3	1	1.6	17	5.7
6. Dissociative reaction, anxiety reaction	1	1.5	4	3.8	3	2.3	8	2.7	12	8.4	1	1.1	1	1.6	14	4.7
7. Conversion reaction	2	2.9	2	1.9	0	0.0	4	1.3	10	7.0	2	2.1	1	1.6	13	4.3
8. Depressive reaction	2	2.9	1	1.0	0	0.0	3	1.0	6	4.2	2	2.1	0	0.0	8	2.7
9. Obsessive-compulsive reaction	4	5.9	1	1.0	1	0.8	6	2.0	2	1.4	0	0.0	0	0.0	2	0.7
10. Phobic reaction	2	2.9	0	0.0	0	0.0	2	0.7	3	2.1	0	0.0	0	0.0	3	1.0
11. Psychophysiological retion	1	1.5	5	4.8	0	0.0	6	2.0	8	5.6	0	0.0	0	0.0	8	2.7
12. Schizophrenic reaction, all types	4	5.9	2	1.9	2	1.6	8	2.7	2	1.4	10	10.6	0	0.0	12	4.0
13. Emotionally unstable personality, inadequate personality	1	1.5	1	1.0	1	0.8	3	1.0	0	0.0	2	2.1	2	3.2	4	1.3
14. Passive-aggressive personality	0	0.0	3	2.9	1	0.8	4	1.3	1	0.7	0	0.0	2	3.2	3	1.0
15. Personality trait disturbance	0	0.0	1	1.0	3	2.3	4	1.3	0	0.0	1	1.1	0	0.0	1	0.3
16. Schizoid personality	3	4.4	1	1.0	0	0.0	4	1.3	2	1.4	1	1.1	0	0.0	3	1.0
17. Sociopathic personality, psychopathic personality	0	0.0	0	0.0	1	0.8	1	0.3	0	0.0	0	0.0	2	3.2	2	0.7
No diagnosis	18	26.5	35	33.7	46	35.9	99	33.0	53	37.1	33	35.1	27	42.9	113	37.7
Totals	68	100.1	104	100.2	128	100.0	300	100.0	143	100.1	94	100.0	63	100.1	300	100.1

REFERENCES

ACHENBACH, T. M. A factor-analytic study of juvenile psychiatric symptoms. Paper read at Society for Research in Child Development, Minneapolis, March 1965.

ACHENBACH, T. M., & ZIGLER, E. Social competence and self-image disparity in psychiatric and nonpsychiatric patients. *Journal of Abnormal and Social Psychology,* 1963, **67,** 197–205.

AMERICAN PSYCHIATRIC ASSOCIATION. *Diagnostic and statistical manual: Mental disorders.* Washington: APA, 1952.

BANDURA, A. Behavioristic psychotherapy. Paper read at School Psychology Institute, University of Wisconsin, July 1964.

BANDURA, A., & WALTERS, R. H. *Adolescent aggression.* New York: Ronald Press, 1959.

BANDURA, A., & WALTERS, R. H. *Social learning and personality development.* New York: Holt, Rinehart, & Winston, 1963.

BARD, J. A., SIDWELL, R. T., & WITTENBROOK, J. M. A practical classification for emotionally disturbed children treated in a welfare setting. *Journal of Nervous and Mental Disease,* 1955, **121,** 568–572.

BENNETT, I. *Delinquent and neurotic children.* New York: Basic Books, 1960.

BURT, C. The unit hierarchy and its properties. *Psychometrika,* **3,** 1938, 151–168.

CHITTENDEN, G. E. An experimental study measuring and modifying assertive behavior in young children. *Monographs of the Society for Research in Child Development*, 1942, **7**, No. 1 (Serial No. 31).

DREGER, R. M. A progress report on a factor-analytic approach to classification in child psychiatry. In R. L. Jenkins & J. O. Cole (Eds.), Diagnostic classification in child psychiatry. *Psychiatric research reports of the American Psychiatric Association*, 1964, **18**, 22–58.

DREGER, R. M., LEWIS, P. M., RICH, T. A., MILLER, K. S., REID, M. P., OVERLADE, D. C., TAFFEL, C., & FLEMMING, E. L. Behavioral classification project. *Journal of Consulting Psychology*, 1964, **28**, 1–13.

EYSENCK, H. J. (Ed.) *Handbook of abnormal psychology*. New York: Basic Books, 1961.

FRUCHTER, B. *Introduction to factor analysis*. Princeton, N. J.: Van Nostrand, 1954.

GLUECK, S., & GLUECK, E. *Unraveling juvenile delinquency*. New York: Commonwealth Fund, 1950.

GLUECK, S., & GLUECK, E. *Predicting delinquency and crime*. Cambridge, Mass.: Harvard University Press, 1959.

GUERTIN, W. H. A factor-analytic study of schizophrenic symptoms. *Journal of Consulting Psychology*, 1952, **16**, 308–312.

HARMAN, H. H. *Modern factor analysis*. Chicago: University of Chicago Press, 1960.

HEWITT, L. E., & JENKINS, R. L. *Fundamental patterns of maladjustment: The dynamics of their origin*. Springfield, Ill.: State of Illinois, 1946.

HOLLINGSHEAD, A. B. Two-factor index of social position. Yale University, 1957. (Mimeo)

HOLLINGSHEAD, A. B., & REDLICH, F. C. *Social class and mental illness*. New York: John Wiley, 1958.

JENKINS, R. L. Diagnoses, dynamics, and treatment in child psychiatry. In R. L. Jenkins & J. O. Cole (Eds.), Diagnostic classification in child psychiatry. *Psychiatric Research Reports of the American Psychiatric Association*, 1964, **18**, 91–120.

LORR, M. The Wittenborn psychiatric syndromes: an oblique rotation. *Journal of Consulting Psychology*, 1957, **21**, 439–444.

LORR, M., KLETT, C. J., & MCNAIR, D. M. *Syndromes of psychosis*. New York: Macmillan, 1963.

MARSHALL, H. R., & MCCANDLESS, B. R. A study in prediction of social behavior of preschool children. *Child Development*, 1957, **28**, 149–159.

MEEHL, P. E. Schizotaxia, schizotypy, schizophrenia. *American Psychologist*, 1962, **17**, 827–838.

PHILLIPS, L., & RABINOVITCH, M. S. Social role and patterns of symptomatic behaviors. *Journal of Abnormal and Social Psychology*, 1958, **57**, 181–186.

PHILLIPS, L., & ZIGLER, E. Social competence: the action-thought parameter and vicariousness in normal and pathological behaviors. *Journal of Abnormal and Social Psychology*, 1961, **63**, 137–146.

RADO, S. Dynamics and classification of disordered behavior. *American Journal of Psychiatry*, 1953, **110**, 406–416.

RIMOLDI, H. J. A. The central intellective factor. *Psychometrika*, 1951, **16**, 75–101.

ROSS, A. O. *The practice of clinical child psychology*. New York: Grune & Stratton, 1959.

UNITED STATES GOVERNMENT. *The dictionary of occupational titles*. Washington: U. S. Government Printing Office, 1949.

WITTENBORN, J. R. Rotational procedures and descriptive inference. *Journal of Consulting Psychology*, 1957, **21**, 445–447.

WITTENBORN, J. R., & HOLZBERG, J. D. The generality of psychiatric syndromes. *Journal of Consulting Psychology*, 1951, **15**, 372–380.

WITTENBORN, J. R., HOLZBERG, J. D., & SIMON, B. Symptom correlates for descriptive diagnosis. *Genetic Psychology Monographs*, 1953, **47**, 237–302.

ZIGLER, E., & PHILLIPS, L. Social effectiveness and symptomatic behaviors. *Journal of Abnormal and Social Psychology*, 1960, **61**, 231–238.

ZIGLER, E., & PHILLIPS, L. Psychiatric diagnosis: A critique. *Journal of Abnormal and Social Psychology*, 1961, **63**, 607–618.

(Received October 25, 1965)

Vol. 80, No. 8 Whole No. 616, 1966

Psychological Monographs: General and Applied

THE INFLUENCE OF BELIEF SYSTEMS ON INTERPERSONAL PREFERENCE: A VALIDATION STUDY OF ROKEACH'S THEORY OF PREJUDICE[1]

DAVID D. STEIN[2]

University of California, Berkeley

A full-scale test of Rokeach's theory of belief prejudice with 630 9th-grade students strongly supports the validity of the theory. When information about a stimulus person's beliefs in the area of personal values is made available, similarity or dissimilarity in beliefs is the primary determinant of attitudes of white gentiles toward Negroes and Jews. These results also hold for Negro and Jewish students' attitudes towards members of the majority. Only secondarily does racial or religious affiliation per se, or high versus low relative socioeconomic status, influence the students' feelings (friendliness measure) and action orientations (social distance scale) toward others. In response to individual social distance items, gentile Ss showed relative unwillingness to interact with Negroes as compared with whites in "sensitive" areas of interracial contact. Similar results, but to a much lesser degree, were obtained for anticipated interaction with Jewish stimulus persons. Gentile Ss' responses on another occasion to an otherwise undescribed "Negro teenager" correlated substantially with their responses to a lower status Negro to whom values unlike their own were ascribed. Other data indicate strong race and religion effects and a weaker status effect in the absence of information about stimulus persons' beliefs.

One of the many ideas presented in *The Open and Closed Mind* (Rokeach, Smith, & Evans, 1960) is that prejudice may be in large part the result of perceived dissimilarity of belief systems. In essence, Rokeach et al. (1960) contended that the prejudiced person does not reject a person of another race, religion, or nationality because of his ethnic membership per se, but rather because he perceives that the other differs from him in important beliefs and values. This specific hypothesis was developed out of Rokeach's general theoretical framework, in which the emphasis is on cognitive determinants of social behavior and belief systems are given focal attention.

Rokeach et al. report two studies in which subjects were asked to rate pairs of stimulus individuals on a 9-point scale, defined at the ends by the statements, "I *can't* see myself being friends with such a person" and "I can *very easily* see myself being friends with such a person." In one experiment, the stimulus individuals were white or Negro; in the other they were Jewish or gentile. Reported beliefs of the stimulus individual concerning racial, religious, and other matters were also varied. It was found that the friendship preferences expressed were determined primarily on the basis of congruence in beliefs rather than on racial or religious grounds.

The presentation of the theory of belief prejudice, based on these preliminary experiments, has led to a number of studies that both lend support to and qualify the basic tenets of the theory (Byrne & Wong, 1962; Rokeach & Mezei, 1966; Stein, Hardyck, & Smith, 1965; Triandis, 1961; Triandis & Davis, 1965). These studies have developed along the following lines:

[1] This paper is based on a doctoral dissertation submitted to the Graduate Division and Psychology Department of the University of California, Berkeley, June 1965. This research was supported by Grant MH 10610-01 from the National Institutes of Health, United States Public Health Service to M. Brewster Smith, principal investigator. The author wishes to express his gratitude to Professor Smith and to Jane Allyn Hardyck for their helpful suggestions and critical comments in the preparation of this paper.

[2] Now at the Department of Psychiatry, Albert Einstein College of Medicine, Yeshiva University.

1. Triandis (1961) objected to the use by Rokeach et al. of a single dependent variable of friendship and instead employed a social distance scale. Varying race, religion, occupational status, and similarity of philosophy of life to that of the subject, Triandis obtained a "race effect" that accounted for about four times as much variance as any of the other three effects singly. Triandis concluded that race, rather than belief, is the primary determinant of prejudice.

2. Rokeach (1961) criticized Triandis' manipulation of similarity of philosophy via Morris' (1956) "13 ways to live," as based on complex and diffuse paragraphs that would not make similarity or difference of beliefs salient to the subjects.

3. Byrne and Wong (1962) essentially supported Rokeach's position, employing alleged responses to an attitude questionnaire as the basis for manipulating similarity of belief, and personal feelings of friendliness and willingness to work together in an experiment as dependent variables.

4. Stein et al. (1965), attempting to reconcile the disparate findings, followed Byrne and Wong in constructing stimulus individuals who were intended to appear more real to their subjects than had been the case in the Rokeach et al. and Triandis studies. Their modification also required subjects to respond to stimulus persons individually rather than in pairs in order to minimize any awareness that the choice was between response in terms of race or of belief. As dependent variables, Stein et al. employed both a measure of friendly feelings and a social distance scale, on which responses to each of the individual items as well as to the total scale could be separately analyzed. Their findings, which provide the starting point for the present study, included the following:

First, in the analysis of "friendliness" responses and total social distance scale scores, belief accounted for much more variance than race, although both effects were significant. Secondly, strong "race effects" were obtained on "sensitive" items in the social distance scale, perhaps reflecting institutionalized areas of prejudice. There were significant race effects, and, to a lesser degree, status effects on the total scores on a social distance scale administered to the same subjects on a previous occasion when race and status had been varied and no information about beliefs provided. Thirdly, a correlational analysis showed that subjects responded to a Negro stimulus person presented as unlike them in values in much the same way as they had previously responded to an otherwise unspecified Negro about whom they had no other information ($r = .62$). This correlation was interpreted to mean that, in the absence of other information, subjects assume that Negroes are unlike them in values. Stein et al. (1965) concluded:

> When subjects are forced to evaluate stimulus individuals in terms of their beliefs, then belief congruence is more important than race. But when the belief component is not provided, spelled out in considerable detail, subjects will react in racial terms on the basis of assumptions concerning the belief systems of others, and of emotional or institutionalized factors [p. 289].

5. Rokeach and Mezei (1966) were interested in seeing if the theory of belief prejudice could be generalized beyond the pencil-and-paper test situations to behavior in representative real life settings. In three interrelated experiments, a naïve subject was asked either to state a preference for two of four confederates to take a coffee break, or to choose among fellow "job applicants" the two with whom he would most like to work. Two of the four confederates were Negro and two were white. One of each race agreed and one disagreed with the subject. The authors conclude that in all three experiments, similarity of belief is the most frequent basis of subjects' choices.

6. Recently, Triandis and Davis (1965) reported a study in which 300 subjects responded on 12 semantic and 15 Behavioral Differential scales to eight stimulus persons generated by all possible combinations of the characteristics Negro-white, male-female, and pro or con civil rights legislation. Some subjects proved to be extremely sensitive to the race of the stimulus persons

while other subjects showed a greater sensitivity to the beliefs of the stimulus persons. The likelihood of a person's showing sensitivity to race as compared to belief is related to the degree of intimacy of the behavioral situations described. For almost all subjects, the more intimate the behaviors, the more frequently did subjects respond in terms of race. For the least intimate behaviors, most subjects responded in terms of belief. When the behaviors were intermediate in intimacy, subjects characterized independently as "racially prejudiced" responded in terms of race, and subjects characterized independently as "belief prejudiced" responded in terms of belief. These findings are generally consistent with those of Stein et al., especially with regard to the importance of race in determining responses to intimate items on a social distance scale. Although Stein et al. did not have an independent measure of their subjects' prejudicial orientation, that is, belief or race, they found, contrary to Triandis and Davis, that belief was equally important throughout the social distance scale.

RATIONALE AND AIMS

The present research was undertaken to replicate the original study by Stein et al. (1965) with a more adequate sample, and to elaborate upon it in a variety of ways. In connection with another study[3] questionnaires had been administered in the Spring of 1963 to the entire eighth grade of the Commutertown[4] public school system. Crucial for the present study was the inclusion of a series of items tapping the respondents' beliefs in the area of personal values. The data to be reported here were collected from the same students in the late Spring of 1964 when they were ninth graders.

Specifically, the present study replicates all of the analyses of white gentile students' attitudes toward Negro stimulus teenagers. By way of extension, it assesses subjects' responses to stimulus teenagers composed so as to vary systematically not only race (white versus Negro) and similarity of belief, but also social status. Further, the generality of findings is extended by having half the sample respond to stimulus adults rather than to stimulus teenagers. In some analyses, the religion rather than race of the stimulus person is varied with belief and status. The larger and more heterogeneous sample in the present study permits sex differences to be examined and separate analyses made for Jewish, Negro, white Protestant, and white Catholic subjects, the former two groups being unrepresented in the earlier study.

The inclusion of religious affiliation as a variable has interesting implications in terms of Rokeach's theory. Knowledge of a person's religion yields information about central features of his probable beliefs. Thus, when both religion and belief (exemplified in personal values) are varied in the presentation of stimulus persons, strong elements of the belief component are embedded in the meanings attached to the religion ascribed. If Rokeach is right, ascription of religion rather than race might be expected to have a large effect, in comparison with similarity of belief. In addition, religious membership should be particularly salient for Jewish subjects, and race for Negro subjects, because of the emphasis on these factors in their upbringing. To pit religion and race, respectively, against similarity of belief, as determinants of these subjects' responses to stimulus persons, is thus to test Rokeach's theory of belief prejudice under quite stringent conditions.

METHOD

Preparation of Questionnaires

Each subject received a personally tailored questionnaire built around supposed excerpts from the replies of four teenagers or adults who had allegedly filled out the same research questionnaire

[3] A large-scale study of adolescent intergroup relations and attitudes being conducted by Jane Allyn Hardyck and M. Brewster Smith through the Survey Research Center, University of California, with the support of the Anti-Defamation League of B'nai B'rith provided the opportunity for the present investigation.

[4] Commutertown is the fictitious name given to a Northeast suburban city. I am indebted to the superintendent and staff of the Commutertown schools, who must remain anonymous, for their cooperation, and to Oscar Cohen of the Anti-Defamation League for his part in securing it.

that the subject himself had completed in the eighth grade. Whether the stimulus persons whom any one subject received were teenagers or adults depended upon the form of the questionnaire that the subject had previously filled out. This original questionnaire contained either questions regarding the subject's feelings about teenagers or parallel questions about adults, and subjects at that time had been randomly assigned to receive *one* form or the other.

The instructions which appeared on the first page of the questionnaire were as follows: (Form A)

> As you will probably remember, about a year ago we asked you to answer some questions concerning your interests and attitudes about yourself, your friends, and certain groups of people. You may also recall that there were some questions asking you to give first impressions about people when you knew only a few things about them, such as the person's religion or type of job. We are very much interested in how people form these impressions.
>
> In fact, we would like to know how you would feel about some adults who took a similar questionnaire to the one you answered, but in other parts of the country. Therefore, we have taken some of their answers and presented them on the following pages.
>
> We want you to look at the descriptions of *four* adults and answer some questions about how you feel toward each of them. It is important that you go through this booklet in order. Do *not* skip ahead, and once you have finished answering questions about a person, do not go back.
>
> If you have any questions, please raise your hand and your teacher will help you. Be sure to read *everything* carefully. And remember, feel free to answer the questions exactly the way you feel, for no one but the research workers at the University of California will see your answers.

The instructions and basic format of the questionnaire follow the plan of the questionnaire used by Stein et al. (1965), with only minor modifications. Stimulus persons were in each case the same sex as the subject. The presentation of each stimulus person contained information about belief, race or religion, and status.

The value items for presenting the belief system of a stimulus adult were as follows:

Do you think people in general *ought* to . . .

1. Be loyal to the U.S. more than to any other group or cause.
2. Be interested in doing things in their community; be useful citizens.
3. Be *un*concerned with making a great deal of money.
4. Be intelligent and well informed, be able to think clearly about things.
5. Keep their property in good condition; not let things get run down.
6. Have good taste in clothes.
7. Be concerned about other people; *not* be self-centered.
8. Be modest, *not* try to draw attention to themselves.
9. Support movements or groups that are working for equal rights for everyone.
10. Be sincerely religious.
11. Have respect for other people's wishes and beliefs; *not* be bossy.
12. Let everyone have his fair share in running business and politics in this country.
13. Be honest and trustworthy.
14. Be generally friendly and sociable; mix with different kinds of people.
15. Treat other people as equals; *not* be conceited or snobbish.
16. Follow all the rules and laws that have been made by those in authority.
17. Stay in groups or neighborhoods where they are welcome; *not* be "social climbers."
18. Live up to strict moral standards.
19. Be hard working, *not* lazy.
20. Go along with what most others do and stand for; *not* be too different.

These items were followed by five columns of response alternatives headed by "Strongly feel they should" to "Strongly feel they shouldn't" with "Don't care" as the middle point. The experimenter circled the appropriate alternative for each item, as designated by the computer program (see p. 5), to give the subject the impression of the stimulus person's responses to these items.

The information, other than values, provided to describe a stimulus adult was as follows: (Form for Religion)

1. Sex________
2. Age________
3. What is your job called?________
4. How much education did you have?
   ______some grade school
   ______finished the 8th grade
   ______some high school
   ______graduated from high school
   ______some college
   ______graduated from college
   ______some further education after college
5. What is your religion?
   ______Protestant ______Catholic
   ______Jewish ______other ______none

A copy of the complete Form T questionnaire appears in Stein (1965). The factors and their corresponding levels are given as follows:

Race: White versus Negro
Religion: Protestant or Catholic versus Jewish
Belief: Similar values versus dissimilar values
Status: High versus low

The four stimulus persons, who varied in terms of the factors of race and religion, were assigned to each subject following the plan presented in Table 1. Each pair of stimulus persons indicated in the table was composed of one high in status and one low in status.

Description of stimulus persons: Belief factor. Each subject had filled out a 20-item value scale concerning "how people ought to be" (Form A) or "how fellow students ought to be" (Form T), as part of the Interest and Attitude questionnaire given when he was in the eighth grade. (For the adult form, see above; for the teenage form, see Stein et al. [1965, p. 283], with the omission of Items 1, 3, 17, 21, and 22 because of small response variance.) An IBM 7090 computer program was written so that each subject's original responses to these items could be presented systematically in such a way as to make two sets that were similar and two that were dissimilar to the subject's original responses. The basic procedure for this program appears in Stein (1965, pp. 94–97; for complete details write to the author). Thus, beliefs were ascribed to the four stimulus persons presented to any given subject on the basis of the subject's own responses to these same items. One of the four stimulus persons was always described with exactly the same responses that the subject originally gave to the items. In order to avoid raising the subject's suspicions, the other like-valued stimulus person was made to differ slightly from the first by changing a few responses one step on the 5-point scale. The two unlike-valued stimulus persons were prepared by making more radical changes, again using the subject's own responses as the reference point. In addition, the program randomly varied the order of the two like-valued patterns and the two unlike-valued patterns.

Description of stimulus persons: Race or religion and status factors. The information about race or religion and status was also presented by checks in the appropriate spaces, as if representing questionnaire responses. Each stimulus person was described in terms of *either* race or religion.

For those subjects who responded to the adult form of the questionnaire, occupation and education were used as indexes of status. A "doctor" or "lawyer," randomly interchanged, was combined with "some further education after college" in the descriptions of high-status males. For low-status males, a "factory worker" and "truck driver" were randomly interchanged, and the amount of education attributed to them was "some grade school." A high-status female was presented as either an "executive secretary" or a "dress designer" with "some further education after college." A low-status female was depicted as a "factory worker" or a "waitress" with an education of "some grade school." Each adult stimulus person was described as either 34 or 36 years old.

All teenage stimulus persons were described as in the ninth grade (same grade as the subjects). Status was indicated by program in school and last year's grade average: "college preparatory" and "getting about a 'B' average" for high status, and "vocational" and "getting below a 'D' average" for low status.

TABLE 1

ASSIGNMENT OF STIMULUS PERSONS TO SUBJECTS

Membership group of subject	Stimulus persons: Jewish	Negro	Protestant	Catholic	White
Jewish	2	—	2	—	—
Negro	—	2	—	—	2
Half of the white Protestants	2	—	2	—	—
Half of the white Protestants	—	2	—	—	2
Half of the white Catholics	2	—	—	2	—
Half of the white Catholics	—	2	—	—	2

Experimental Design

Given the variables with which we are concerned, eight possible stimulus persons could be constructed. Excessive time demands (as indicated by a pilot study) made it impractical for each subject to respond to all eight combinations. Therefore, a $2 \times 2 \times 2$ factorial design in blocks of Size 4 (repeated measures) was employed (Winer, 1962, pp. 409–412). According to this design, the comparisons involving race, for example, are as follows:

One-half of the subjects received the following four stimulus persons:

Group I
White, unlike values, lower status
Negro, like values, lower status
Negro, unlike values, upper status
White, like values, upper status

One half of the subjects received the following four stimulus persons:

Group II
Negro, unlike values, lower status
White, like values, lower status
White, unlike values, upper status
Negro, like values, upper status

In the comparisons involving religion, Jewish and Protestant or Catholic were substituted for Negro and white, respectively. (See Table 1; note in the religion comparisons that "same versus different" religion is the basis for ascribing religion to the stimulus persons. Thus, Catholic subjects responded to Catholic and Jewish stimulus persons and Protestant subjects responded to Protestant and Jewish stimulus persons.)

Within each subsample (the 24 cells in Table 2), each subject was randomly assigned to Group I or II above and the order of presentation of stimulus persons within both Groups I and II was randomly varied, with the restriction that no two "like-valued" or "unlike-valued" stimulus persons ever appeared consecutively in a questionnaire.

Dependent variables. After the description of

TABLE 2

SAMPLE ($N = 630$)

	Subjects											
	Jews		Negroes		Protestant				Catholic			
Form of questionnaire	Adult	Teenage	Adult	Teenage	Adult	Adult	Teenage	Teenage	Adult	Adult	Teenage	Teenage
Boys	69	84	25	23	12	15	10	7	18	17	16	13
Girls	68	88	25	24	15	13	8	8	18	17	18	19
Stimulus persons	Jewish and gentile	Jewish and gentile	Negro and white	Negro and white	Negro and white	Jewish and gentile	Negro and white	Jewish and gentile	Negro and white	Jewish and gentile	Negro and white	Jewish and gentile

each stimulus person, the following sets of questions were asked:

1. Friendliness. The first question, which measured the subject's overall affective reaction to the stimulus person and which might be considered a measure of the affective component of a prejudiced or unprejudiced attitude, appeared as follows:

> If you met this person for the first time, what would your immediate reaction be?
>
> I think I would feel:
>
> _____quite friendly
> _____a little friendly
> _____nothing either way
> _____a little unfriendly
> _____quite unfriendly

2. Social distance. Next, subjects responded to a 10-item social distance scale. The items on the scale are appropriately designed for the teenage form as well as the adult form. (See Stein et al., 1965, p. 287, for the complete wording of Form T items or Table 11 hereafter for a paraphrased version.)

The items on the adult form were as follows:

Do you think you would be willing to . . .

Yes No

____ ____ have this person as a neighbor on your street
____ ____ work on a charity fund-raising drive with this person
____ ____ have this person as one of your speaking acquantances
____ ____ go to a party to which this person was invited
____ ____ have this person as a member of your social group or club
____ ____ live in the same apartment house as this person
____ ____ have this person as a close personal friend
____ ____ invite this person to your home for dinner
____ ____ have a close relative marry this person
____ ____ share an apartment with this person

3. Check on the manipulation. Thirdly, a question was asked regarding how much like the stimulus person the subject saw himself. This question appears as follows:

> How much like you would you say this person is?
>
> _____as much like me as any person I can think of
> _____very much like me
> _____a little like me
> _____a little unlike me
> _____very much unlike me
> _____as much unlike me as any person I can think of

Since the stimulus persons were designed to be "like" or "unlike" the subject, it was necessary to find out if subjects so perceived them.

Sample

The final sample consisted of 630 ninth-grade students. Table 2 classifies the subjects by ethnic

(racial or religious) affiliation, sex, and form of the questionnaire, and by the particular stimulus persons who appeared in the subjects' questionnaires. In order to obtain sufficiently large cell entries, subjects were combined across the two junior high schools and Protestants and Catholics were combined (as "gentiles") for most of the analyses.

Administration of Questionnaires

The questionnaires were administered during the subjects' regular class in social studies. The investigator met with the social studies faculty of each school the day before the administration to go over the instructions and to answer any questions. Seven social studies teachers at one of the junior high schools and nine at the other administered the procedure.

Each subject received his questionnaire in a sealed envelope. His name appeared on the envelope but his questionnaire was identified only by a code number. The teachers were told to throw away the envelopes as soon as the students had removed their questionnaires. Subjects were told that the code number was necessary for statistical analyses, that no student would be considered individually or by name, and that only the research workers at the University of California would see their answers.

Results and Discussion

The $2 \times 2 \times 2$ factorial design used for the analysis of the check on the manipulation and for the two dependent variables of liking and social distance can demonstrate only whether or not a given independent variable has a significant effect on the dependent variable. In order to determine whether one treatment effect is significantly greater than another, it is necessary to calculate the proportion of total variance contributed by each treatment effect and then to test for any significant difference between effects.

The index, Ω^2, expresses the strength of association between independent and dependent variables in terms of the proportion of total variance accounted for by the treatment effect (Hays, 1963). The Ω^2 values were computed for each sample and for each treatment effect in the present study. The Ω^2 values for any two selected factors were then ranked across samples in order of magnitude and White's rank test (Edwards, 1954) was applied to determine whether one factor contributed a significantly greater proportion of the variance than the other. Two-tailed tests were used in all cases.

Throughout the discussion, trends in the differential effect of treatments on related subsamples will be pointed out. Quite often, though, significance tests could not be meaningfully applied to these data since they were based on so few samples.

Check on the Manipulation

The subjects' responses to the question: "How much like you would you say this person is?" served as a check as to whether the stimulus persons appeared like or unlike the subjects as intended. A summary of the 16 analyses of variance for responses to this question appears in Table 3. The main effect of belief accounts for almost all of the variance, contributing significantly more variance than either the race, the religion, or the status effect ($p < .01$ for all rank-order comparisons). Subjects tended to see themselves as similar or dissimilar to the stimulus persons mainly in terms of belief. Means and standard deviations for many of these analyses are omitted here to conserve space, but are presented in Stein (1965). In this analysis, like-valued persons are perceived as more like the subject than unlike-valued ones.

In the analysis involving the religion variable, the small proportion of variance not accounted for by the belief effect is divided about equally between status and religion. The fact that status effects account for more variance on Form T than on Form A ($p < .05$ for the rank-order difference) suggests that the status attributes of the stimulus teenagers are more salient to the subjects than the corresponding status attributes of the adult stimulus persons. The two Form A samples for which the status effect was significant are the gentile and Jewish males. This finding seems reasonable since the manipulation of the adult stimulus persons' status may well have been more powerful for males than for females. That is, the difference between "doctor" and "lawyer" on the one hand and between "factory worker" and "truck driver" on the other appears to be greater than that between "dress designer" and "executive secretary" as opposed to "factory worker" and "waitress."

The fact that three of the four Jewish

samples show significant status effects seems in line with the probable stress on this factor in Jewish families. It is common knowledge that middle-class Jewish parents tend to hope that their children will get good grades, go to college, and enter into professions. The high-status stimulus persons are presented as successful in fulfilling such expectations.

The results concerning the religion variable have some interesting implications. In four of the eight relevant samples there was a significant religion effect; whereas in no sample was there a significant race effect ($p < .05$ for the rank-order difference between Ω^2 values for religion and race). This fact is in accord with the expectation that there is a meaningful belief component in the ascription of religion. Knowing merely that a person is Protestant, Catholic, or Jewish may imply much about the person's beliefs. Thus it is not surprising that judgments of similarity are more frequently based at least in part on the religion of the subject than on his race, which implies much less about belief systems.

Friendliness

Table 4 shows the summary of the analyses of variance computed on the responses of the 16 subsamples to the "friendliness" question, intended as a measure of "affect" toward the stimulus person.

As may be seen in the column headed Belief, the belief component of the stimulus individuals accounted for almost all of the variance in responses to this question. The belief effect contributes significantly more variance than either the race, the religion, or the status effects ($p < .01$ for all rank-order comparisons). In 15 of the 16 samples there was a highly significant be-

TABLE 3

SUMMARY OF THE 16 ANALYSES OF VARIANCE FOR RESPONSES TO THE QUESTION: HOW MUCH LIKE YOU WOULD YOU SAY THIS PERSON IS? (AS MUCH LIKE ME—AS MUCH UNLIKE ME—6-POINT SCALE)

Sample	Form	N	Race		Belief		Status		Race × Belief	Race × Status	Belief × Status
			F	Prop. of variance	F	Prop. of variance	F	Prop. of variance	F	F	F
Negro males	A	25	.26	.00	17.85****	.10	1.79	.00	.01	.01	.01
Negro females	A	25	.98	.00	66.37****	.39	.10	.00	.58	.01	.01
Gentile males	A	30	1.04	.00	15.26****	.08	4.15*	.02	4.88*	2.88	.72
Gentile females	A	33	.28	.00	135.19****	.47	.03	.00	.03	1.12	.28
Negro males	T	23	.56	.00	31.62****	.21	2.25	.01	.88	4.26	.88
Negro females	T	24	.12	.00	68.84****	.40	7.98***	.04	1.12	.28	1.12
Gentile males	T	26	.12	.00	35.80****	.20	7.93***	.04	.03	2.51	.77
Gentile females	T	26	.08	.00	64.08****	.34	7.12***	.03	1.43	.68	1.43
			Religion		Belief		Status		Religion × Belief	Religion × Status	Belief × Status
Jewish males	A	69	15.82****	.03	148.83****	.28	18.01****	.03	1.42	2.13	.44
Jewish females	A	68	3.08	.00	350.25****	.54	1.05	.00	.02	.08	1.37
Gentile males	A	32	5.15*	.02	48.59****	.24	.24	.00	1.68	.42	.24
Gentile females	A	30	.17	.00	121.88****	.41	2.13	.00	1.08	.04	1.56
Jewish males	T	84	9.20***	.01	151.35****	.26	70.75****	.12	0.0	.40	4.98*
Jewish females	T	88	.24	.00	350.10****	.42	64.02****	.08	.74	.06	10.24***
Gentile males	T	20	3.60	.03	.99	.00	.81	.00	.81	0.0	.88
Gentile females	T	27	9.73***	.05	31.54****	.18	5.79*	.03	.07	.64	2.30

* $p = .05$.
*** $p = .01$.
**** $p = .001$.

TABLE 4

SUMMARY OF THE 16 ANALYSES OF VARIANCE FOR RESPONSES TO THE QUESTION: IF YOU MET THIS PERSON FOR THE FIRST TIME, WHAT WOULD YOUR IMMEDIATE REACTION BE? (QUITE FRIENDLY—QUITE UNFRIENDLY—5-POINT SCALE)

Sample	Form	N	Race F	Race Prop. of variance	Belief F	Belief Prop. of variance	Status F	Status Prop. of variance	Race × Belief F	Race × Status F	Belief × Status F
Negro males	A	25	2.67	.01	21.96****	.14	.30	.01	.30	.96	.11
Negro females	A	25	3.54	.02	55.34****	.34	.04	.00	.81	.20	.65
Gentile males	A	30	2.06	.00	20.50****	.10	5.38*	.02	5.38*	.01	.11
Gentile females	A	33	.50	.00	80.31****	.34	.82	.00	.25	2.28	.50
Negro males	T	23	.25	.00	23.75****	.17	.01	.00	1.20	2.22	.80
Negro females	T	24	.01	.00	41.54****	.27	.97	.00	2.02	.01	1.44
Gentile males	T	26	.48	.00	38.56****	.21	6.40*	.03	0.0	.85	0.0
Gentile females	T	26	.18	.00	42.58****	.25	4.43*	.02	1.60	.04	.71
			Religion F	Religion Prop. of variance	Belief F	Belief Prop. of variance	Status F	Status Prop. of variance	Religion × Belief F	Religion × Status F	Belief × Status F
Jewish males	A	69	10.43***	.02	116.41****	.25	5.78*	.01	1.06	1.36	.57
Jewish females	A	68	1.14	.00	224.59****	.42	1.14	.00	.37	.02	.09
Gentile males	A	32	2.23	.01	26.16****	.17	.15	.00	.15	2.23	.50
Gentile females	A	30	.22	.00	78.92****	.34	.22	.00	.87	1.97	.22
Jewish males	T	84	2.66	.00	113.21****	.22	25.26****	.05	0.0	.96	.96
Jewish females	T	88	.02	.00	229.85****	.36	20.80****	.03	.74	1.23	.74
Gentile males	T	20	.80	.00	9.19****	.08	1.00	.00	7.49***	.62	.02
Gentile females	T	27	3.46	.01	44.56****	.24	5.28*	.02	.97	.59	4.32*

* $p = .05$.
*** $p = .01$.
**** $p = .001$.

lief effect ($p < .001$), and the belief effect in the other sample (gentile males, Form T) was significant at better than the .01 level. (Note that this is the same sample for which the manipulation of similarity appeared to be ineffective.)

The results for this sample in all analyses should be viewed with great caution. Of the 16 subsamples in the study, this 1 has the smallest N (20). Besides, when this sample was divided so that approximately half of the subjects would receive four treatments and the other half of the subjects the other four treatments, the actual split came out to be 13 and 7 instead of the desired 10 and 10 because some subjects were absent or had transferred to another school. The Winer (1962) model assumes an equal number of subjects in each group. If N is large and the difference between N_1 and N_2 in each group is small, statistical assumptions for the model are not seriously violated. Departures from this rule, as in this case, reduce the power of any statistical test that might be applied to the data. The use of a General Linear Hypothesis Model (Biomedical Computer Programs, 1961) is recommended in such cases and was in fact carried out for this sample as well as for the Negro females, Form A ($N = 25$; $N_1 = 15$, $N_2 = 10$). The F ratios reported in Tables 3, 4, and 9 for these samples were derived by this procedure.

The race effect was not significant in any of the eight samples in which race was varied. The religion effect was significant in only one of the eight samples in which it was tested (Jewish males, Form A). The status effect was significant in 7 out of 16 samples, with generally lower significance levels than those for belief. That is, five of these seven tests for status were significant

only at the .05 level while the other two reached the .001 level. Moreover, rank-order tests between the amount of variance explained by these three factors showed no significant differences.

There are no apparent sex differences on the status factor. Five of the seven samples that showed significant status effects had responded to Form T—a trend paralleling findings previously reported in regard to the effectiveness of the manipulations. Three out of the four Jewish samples show significant status effects although very little variance is contributed by these samples. Only 3 of the 48 tests for two-way interactions were significant, a finding that could easily have arisen by chance, especially since there is no correspondence between the groups showing such interaction effects in this table and in Table 3 concerning the check on the manipulation.

These findings lend strong support to Rokeach's theory. Subjects' affective responses to stimulus persons are much more strongly influenced by ascribed similarity of belief systems than by ascribed religion or race.

Correlational Analysis of Responses to the Friendliness Item

Adult Negro stimulus persons. Essentially the same "friendliness" question had also been asked, in a somewhat different format, on the Interest and Attitude questionnaire, given to the same subjects a year before the present study. At that time, subjects had been asked to respond to a list of many different categories of persons, of which two were "a Negro" (or "a Negro teenager") and "a Jew" (or "a Jewish teenager"). It had been originally hypothesized (Stein et al., 1965, p. 286) that subjects' responses to the Negro teenager should correlate moderately both with responses to a like-valued Negro and an unlike-valued Negro unless, for some reason, subjects have an expectation that Negro teenagers in general are either like them or unlike them in values. The finding that responses to "a Negro teenager" correlated highly with responses to a Negro teenager with unlike values but not with responses to a Negro with like values was of considerable import. Therefore, a similar analysis was carried out in the present study.

Of the 630 subjects, 572 had answered the appropriate items in the previous questionnaire. The 572 subjects represent 16 subsamples, of which 4 each responded to "a Negro," "a Negro teenager," "a Jew," and "a Jewish teenager."

In the study by Stein et al., mean responses to "a Negro teenager" fell between the means for the other two experimentally presented stimulus teenagers. It seemed reasonable that subjects should feel friendliest toward a Negro with like values, followed by the unspecified "Negro teenager," and finally by the least preferred stimulus teenager, the Negro with unlike values. One might expect similar outcomes among the subsamples in the present study unless the subjects are affected by the addition of status as a variable rather than as a constant in the description of like- and unlike-valued stimulus persons, or by the fact that stimulus adults elicit different responses than stimulus teenagers, or unless Negro and Jewish subjects respond differently from white gentile subjects. A further difference is the fact that a year intervened between administrations of the questionnaires in the present study, as compared with only 2 months in the former one.

Table 5 summarizes the results for the subjects in the four samples who responded to "a Negro" (Form A). In the column headed $\bar{X}$, the first six means reflect the responses of Negro subjects who received Form A. For each of these two samples "a Negro" is the most preferred stimulus person, although these means are not significantly different from those for the like-valued Negro of either upper or lower status. Since these are Negro subjects, and this score was taken from the previous questionnaire when the "Negro" was embedded in a list of other stimulus persons, the salience of the Negro stimulus was apparently increased and the subjects responded quite favorably. Means for "a Negro" and the like-valued Negro are both significantly different from the means for the unlike-valued Negro. These results, therefore, are consistent with find-

TABLE 5

ANALYSIS OF RESPONSES ON THE "FRIENDLINESS" ITEM TOWARDS NEGRO ADULT STIMULUS PERSONS

Sample[a]	N	Stimulus persons	$\bar{X}$[c]	σ^2	Comparison	r	t between means	t between σ^2
Negro	23	Negro-unlike values—lower status	2.4	1.69	Negro-unlike values—lower status Negro-like values—upper status	−.15	2.66**	3.05***
		Negro-like values—upper status	1.5	.49	Negro-unlike values—lower status A Negro[b]	.17	3.32***	3.06***
		A Negro[b]	1.4	.49	Negro-like values—upper status A Negro[b]	.41	0.53	0.0
Negro	23	Negro-like values—lower status	1.4	.36	Negro-like values—lower status Negro-unlike values—upper status	.14	4.14***	2.47*
		Negro-unlike values—upper status	2.4	1.0	Negro-like values—lower status A Negro[b]	−.05	0.44	0.0
		A Negro[b]	1.3	.36	Negro-unlike values—upper status A Negro[b]	−.08	4.09****	2.45*
Gentile	28	Negro-unlike values—lower status	2.9	.81	Negro-unlike values—lower status Negro-like values—upper status	.50***	4.53***	0.0
		Negro-like values—upper status	2.1	.81	Negro-unlike values—lower status A Negro[b]	−.10	2.54**	1.04
		A Negro[b]	2.2	1.21	Negro-like values—upper status A Negro[b]	−.28	0.23	1.07
Gentile	28	Negro-like values lower status	1.9	.81	Negro-like values—lower status Negro-unlike values upper status	−.41*	1.95	1.13
		Negro-unlike values—upper status	2.6	1.21	Negro-like values—lower status A Negro[b]	.22	0.82	0.0
		A Negro[b]	2.1	.81	Negro-unlike values—upper status A Negro	.05	1.72	1.03

[a] Each sample involves different subjects: boys and girls are combined.
[b] From questionnaire given when students were in the eighth grade.
[c] A low score indicates greater friendliness toward the stimulus person.
* $p < .05$.
** $p < .02$.
*** $p < .01$.
**** $p < .001$.

ings in the former study. However, none of the correlations between responses to the possible combinations of stimulus persons is significant.

The final six means in the $\bar{X}$ column in Table 5 represent responses of the two "gentile" samples. Here the means are ordered as predicted, but responses to the unlike-valued Negro stimulus persons are significantly different from those to the other two stimulus persons for only the first gentile sample. In addition, two significant correlations emerge, neither of which was predicted. A correlation of .50 ($p < .01$) between an unlike-valued Negro of lower status and a like-valued Negro of upper status does not make any apparent sense. Likewise in the other gentile sample, a correlation of $-.41$ ($p < .05$) between a like-valued Negro of lower status and an unlike-valued Negro of upper status is not readily interpretable.

Teenage Negro stimulus persons. Table 6 presents the results for subjects in the four samples who responded to a "Negro teenager" in addition to the experimentally presented Negro teenage stimulus persons. The two gentile samples in Table 6 provide a replication of the corresponding analysis in the study by Stein et al. (1965). The first six means under the column headed $\bar{X}$ represent the responses of two Negro samples. These subjects again show greatest friendliness toward "a Negro teenager" rather than to the like-valued Negro; the only mean difference that is *not* significant is between "a Negro teenager" and "Negro–like values—upper status." These results, then, are essentially consistent with the findings for the Negro subjects who took Form A. The one significant correlation obtained for these two Negro samples is .52 ($p < .02$) between responses to "a Negro teenager" and to a like-valued Negro of lower status. Although this result was not specifically predicted, it seems to make sense. Negro subjects responded to "a Negro teenager" in much the same way as they did to a Negro who is like them in values and has lower status. Since the Negro subjects themselves come mainly from lower class families, the attributes of this stimulus person actually resemble their own most closely.

The last six means under the column $\bar{X}$ in Table 6 present the scores for the two gentile samples who responded to "a Negro teenager." For both samples, the means are ordered as predicted, although in the second sample, the unlike-valued Negro is not significantly different from the "Negro teenager." It seems remarkable that these means order as predicted in almost all samples considering that a year separated the responses to "a Negro teenager," and the other stimulus individuals. The only significant correlation that emerges is one of .53 ($p < .01$) between responses to "a Negro teenager" and an unlike-valued Negro of lower status, a result in good accord with the rationale underlying the earlier findings, given the addition of status as a new variable, and the substantial exposure of the present samples (unlike the sample of the previous study) to Negroes, who were predominantly from lower class backgrounds.

The fact that significant interpretable correlations were obtained with Form T but not with Form A suggests that both Negro and gentile samples found it more meaningful to respond to stimulus teenagers than to stimulus adults.

Adult Jewish stimulus persons. A similar correlational analysis was carried out for responses to the Jewish stimulus persons. In the results for Form A (Table 7), the first six means under the column headed $\bar{X}$ are the responses for two Jewish samples. There are significant mean differences between responses to "a Jew" and a like-valued Jew of either upper or lower status, on the one hand, and to an unlike-valued Jew of upper or lower status, on the other. The only correlation that is significant is one between "a Jew" and a Jew with unlike values and upper status ($r = .30$, $p < .01$). This finding is unexpected, and no explanation is offered.

The mean responses for the two gentile samples (last six means of Table 7) are in the order predicted, with greatest friendliness exhibited toward a like-valued Jew, followed by "a Jew," and finally by

TABLE 6

ANALYSIS OF RESPONSES ON THE "FRIENDLINESS" ITEM TOWARDS NEGRO TEENAGER STIMULUS PERSONS

Sample[a]	N	Stimulus person	$\bar{X}$[c]	σ^2	Comparison	r	t between means	t between σ^2
Negro	19	Negro-unlike values—lower status	3.1	1.44	Negro-unlike values—lower status / Negro-like values—upper status	.32	6.34****	3.26***
		Negro-like values—upper status	1.4	.36	Negro-unlike values—lower status / A Negro teenager[b]	.26	6.35****	3.20***
		A Negro teenager[b]	1.3	.36	Negro-like values—upper status / A Negro teenager[b]	.45	0.37	0.0
Negro	23	Negro-like values—lower status	1.8	1.0	Negro-like values—lower status / Negro-unlike values—upper status	.20	3.62***	1.60
		Negro-unlike values—upper status	3.0	1.96	Negro-like values—lower status / A Negro teenager[b]	.52**	3.21***	4.01****
		A Negro teenager[b]	1.3	.25	Negro-unlike values—upper status / A Negro teenager[b]	.06	5.62****	5.61****
Gentile	23	Negro-unlike values—lower status	3.1	1.44	Negro-unlike values—lower status / Negro-like values—upper status	.24	6.08****	3.54****
		Negro-like values—upper status	1.6	.36	Negro-unlike values—lower status / A Negro teenager[b]	.53***	2.42*	0.47
		A Negro teenager[b]	2.6	1.21	Negro-like values—upper status / A Negro teenager[b]	.00	3.54***	2.95
Gentile	27	Negro-like values—lower status	2.2	1.0	Negro-like values—lower status / Negro-unlike values—upper status	.32	2.10*	0.0
		Negro-unlike values—upper status	2.7	1.0	Negro-like values—lower status / A Negro teenager[b]	.23	1.66	0.0
		A Negro teenager[b]	2.6	1.0	Negro-unlike values—upper status / A Negro teenager[b]	.28	0.23	0.0

[a] Each sample involves different subjects; boys and girls are combined.
[b] From questionnaire given when students were in the eighth grade.
[c] A low score indicates greater friendliness toward the stimulus person.
* $p < .05$.
** $p < .02$.
*** $p < .01$.
**** $p < .001$.

TABLE 7

ANALYSIS OF RESPONSES ON THE "FRIENDLINESS" ITEM TOWARDS JEWISH ADULT STIMULUS PERSONS

Sample[a]	N	Stimulus person	$\bar{X}$[c]	σ^2	Comparison	r	t between means	t between σ^2
Jewish	56	Jewish-unlike values—lower status	2.8	.81	Jewish-unlike values—lower status / Jewish-like values—upper status	−.13	7.63****	1.88
		Jewish-like values—upper status	1.6	.49	Jewish-unlike values—lower status / A Jew[b]	.09	8.06****	1.87
		A Jew[b]	1.6	.49	Jewish-like values—upper status / A Jew[b]	.09	0.27	0.0
Jewish	63	Jewish-like values—lower status	1.7	.49	Jewish-like values—lower status / Jewish-unlike values—upper status	−.16	5.48****	3.70****
		Jewish-unlike values—upper status	2.7	1.21	Jewish-like values—lower status / A Jew[b]	.08	0.0	1.05
		A Jew[b]	1.7	.64	Jewish-unlike values—upper status / A Jew[b]	.30***	6.74	2.65**
Gentile	31	Jewish-unlike values—lower status	2.7	.64	Jewish-unlike values—lower status / Jewish-like values—upper status	.19	6.34****	0.74
		Jewish-like values—upper status	1.5	.49	Jewish-unlike values—lower status / A Jew[b]	−.08	2.84****	1.22
		A Jew[b]	2.0	1.0	Jewish-like values—upper status / A Jew[b]	.29	2.45*	2.05*
Gentile	24	Jewish-like values—lower status	1.9	.64	Jewish-like values—lower status / Jewish-unlike values—upper status	.09	5.29****	1.52
		Jewish-unlike values—upper status	3.3	1.21	Jewish-like values—lower status / A Jew[b]	.70***	1.77	2.13*
		A Jew[b]	2.2	1.21	Jewish-unlike values—upper status / A Jew[b]	−.18	3.29***	0.0

[a] Each sample involves different subjects; boys and girls are combined.
[b] From questionnaire given when students were in the eighth grade.
[c] A low score indicates greater friendliness toward the stimulus person.
* $p < .05$.
** $p < .02$.
*** $p < .01$.
**** $p < .001$.

an unlike-valued Jew. A sizable correlation of .70 ($p < .001$) occurs between "a Jew" and a Jew with like values and lower status. Again, no explanation is offered for these unpredicted results—unexpected because the Jewish students in this school system come from predominantly upper-middle-class backgrounds.

Teenage Jewish stimulus persons. The analysis of responses to the Jewish teenager stimulus persons appears in Table 8. For the first six means, "a Jewish teenager" receives the most friendly responses for the two Jewish samples. This finding parallels a corresponding result for the Negro sample and can probably be accounted for by the salience to Jewish students of the Jewish stimulus as embedded in the list of other categories of persons presented for reaction in the original Teenage Interest and Attitude Questionnaire. Only in the second sample are responses to "a Jewish teenager" significantly different from those to the like-valued Jew. But in both samples, responses to the unlike-valued Jew differ significantly from those to both of the other two stimulus teenagers, a finding similar to those obtained in the other samples. No significant correlations appear. The two gentile samples have means that follow the expected order, but in the second sample the only significant difference is that between the Jew with like values and lower status and the unlike-valued Jew. A correlation of .41 ($p < .05$) occurs between responses to the unlike-valued Jew of lower status and the like-valued Jew of upper status. This correlation is almost the same as that for the equivalent pair of Negro adult stimuli. Again, no explanation is offered to account for it.

The correlational data for the Jewish stimulus persons fail to confirm the predictions generally verified in the analyses of the data concerning Negro stimulus persons. Gentile subjects are more prone to express friendliness towards otherwise undescribed Jewish stimulus persons than toward Negro stimulus persons. The interpretation by Stein et al. (1965) with regard to assumed dissimilar belief systems is thus confirmed only for the single analysis that exactly replicates the former study—that for teenager Negro stimulus persons; gentile subjects tend to respond to an otherwise undescribed Negro teenager in the same manner as they do to a Negro who is unlike them in values and has lower status.

Total Social Distance Scale

Among our measures the 10-item social distance scales are probably the best indicators of the subjects' willingness to engage socially in real-life situations with persons similar to the stimulus persons. Total scores on the scales were obtained by summing responses to the ten items, each scored 1 for "Yes" and 0 for "No." Scalogram analyses with other data showed that the social distance scales form very highly reproducible Guttman scales. If the subject omitted no more than three responses to a scale, a "Yes" or "No" response was randomly assigned to each omitted item to facilitate computer analyses. Sixty-one subjects failed to answer enough of the questions basic to the present study and were therefore deleted from the analysis. A separate analysis of the individual items of the social distance scales will be presented in the next section.

A summary of the 16 analyses of variance for responses to the Total Social Distance scales appears in Table 9. In the column headed Belief, it can be seen that similarity of values again accounts for the greatest proportion of variance, accounting for significantly more variance than either the race, the religion, or the status effect ($p < .01$ for all rank-order differences). The belief effect is significant in 15 of the 16 samples at $p \leq .001$; in the remaining sample (gentile males, Form T, Religious Comparison) the belief effect is significant at $p < .05$. For the samples of all three ethnic groups, the belief effect accounts for significantly more variance among girls than among boys (rank-order difference at $p < .01$).

For the first time in the analyses reported here, race appears to have a systematic influence on subjects' responses, although the amount of variance controlled by the race effect is small. Race was var-

TABLE 8

ANALYSIS OF RESPONSES ON THE "FRIENDLINESS" ITEM TOWARDS JEWISH TEENAGER STIMULUS PERSONS

Sample[a]	N	Stimulus person	$\bar{X}$[c]	σ^2	Comparison	r	t between means	t between σ^2
Jewish	81	Jewish-unlike values—lower status	3.1	.81	Jewish-unlike values—lower status / Jewish-like values—upper status	−.13	11.26****	2.28*
		Jewish-like values—upper status	1.5	.49	Jewish-unlike values—lower status / A Jewish teenager[b]	.12	13.76****	3.73****
		A Jewish teenager[b]	1.4	.36	Jewish-like values—upper status / A Jewish teenager[b]	.18	0.51	1.40
Jewish	78	Jewish-like values—lower status	2.0	.81	Jewish-like values—lower status / Jewish-unlike values—upper status	−.04	5.55****	1.70
		Jewish-unlike values—upper status	2.9	1.21	Jewish-like values—lower status / A Jewish teenager[b]	−.08	2.72***	1.03
		A Jewish teenager[b]	1.6	.64	Jewish-unlike values—upper status / A Jewish teenager[b]	.12	8.93****	2.85***
Gentile	26	Jewish-unlike values—lower status	3.2	1.0	Jewish-unlike values—lower status / Jewish-like values—upper status	.41*	7.02****	1.96
		Jewish-like values—upper status	1.8	.49	Jewish-unlike values—lower status / A Jewish teenager[b]	.03	3.01***	0.0
		A Jewish teenager[b]	2.3	1.0	Jewish-like values—upper status / A Jewish teenager[b]	.28	2.31*	1.86
Gentile	19	Jewish-like values—lower status	1.8	.36	Jewish-like values—lower status / Jewish-unlike values—upper status	−.18	2.31*	2.24*
		Jewish-unlike values—upper status	2.5	1.0	Jewish-like values—lower status / A Jewish teenager[b]	.29	1.30	2.30*
		A Jewish teenager[b]	2.1	1.0	Jewish-unlike values—upper status / A Jewish teenager[b]	−.05	1.07	0.0

[a] Each sample involves different subjects; boys and girls are combined.
[b] From questionnaire given when students were in the eighth grade.
[c] A low score indicates greater friendliness toward the stimulus person.
* $p < .05$.
** $p < .02$.
*** $p < .01$.
**** $p < .001$.

ied in eight samples, in four of which there were significant race effects. Three of these four were gentile samples and the other was Negro females, Form A. An examination of the mean scores on the Total Social Distance scale to each of the eight treatments shows that this Negro sample favors Negro to white stimulus persons. The three gentile samples favor white to Negro stimulus persons. Since the social distance scale is designed to assess subjects' commitment to interact with the stimulus persons rather than just feel friendly or unfriendly toward them, race might be expected to play a more important role here than in the case of the "friendliness" item.

As can be seen in the column headed Religion in Table 9, significant effects for religion appear in five of the eight samples in which religion was varied. In four of these five, the effect is significant at only the .05 level and in the other, at the .01 level. It is apparent, however, that the religion effect, like the race effect, contributes a negligible proportion of variance. Nonetheless, religious affiliation becomes important when behavioral commitment rather than diffuse expression of friendliness is involved. The appearance of significant religion effects does not depend systematically on the sex of the respondent or on the Form (A or T) of the questionnaire administered. Note, however, that three of the five samples that showed significant effects for religion were Jewish subjects. It is somewhat reasonable to say that religious membership is particularly salient for Jews, both because of their "minority" status and because of the emphasis on the Jewish way of life in most Jewish homes. This fact may also in part explain why Jewish samples showed particularly strong belief effects; many of the items in the

TABLE 9

SUMMARY OF THE 16 ANALYSES OF VARIANCE FOR RESPONSES TO THE TOTAL SOCIAL DISTANCE SCALE

Sample	Form	N	Race		Belief		Status		Race × Belief	Race × Status	Belief × Status
			F	Prop. of variance	F	Prop. of variance	F	Prop. of variance	F	F	F
Negro males	A	25	2.19	.01	24.49****	.13	.95	.00	.14	.14	0.0
Negro females	A	25	11.67***	.03	126.30****	.33	.04	.00	6.63*	.71	7.96***
Gentile males	A	30	26.42****	.12	16.20****	.07	12.58****	.05	4.46*	2.17	.84
Gentile females	A	33	1.15	.00	88.34****	.30	.31	.00	3.73	2.80	1.34
Negro males	T	23	.18	.00	21.51****	.15	3.24	.02	.42	3.53	2.98
Negro females	T	24	.73	.00	51.36****	.30	1.97	.01	1.02	1.75	1.96
Gentile males	T	26	4.37*	.01	44.64****	.19	17.92****	.07	0.0	5.75*	2.83
Gentile females	T	26	9.07***	.04	37.32****	.20	4.63*	.02	2.67	.90	17.77****
			Religion		Belief		Status		Religion × Belief	Religion × Status	Belief × Status
Jewish males	A	69	6.51*	.01	123.95****	.22	25.12****	.04	.40	.56	.32
Jewish females	A	68	5.48*	.01	236.15****	.35	1.16	.00	.48	.48	.04
Gentile males	A	32	5.38*	.02	61.02****	.28	.03	.00	.06	.80	.15
Gentile females	A	30	0.0	.00	65.63****	.26	3.37	.01	2.74	.16	.05
Jewish males	T	84	5.27*	.01	116.84****	.19	55.59****	.09	.72	2.16	2.16
Jewish females	T	88	1.17	.00	215.93****	.30	56.67****	.08	.60	3.44	7.75***
Gentile males	T	20	.98	.00	6.27*	.06	.23	.00	2.74	.36	.20
Gentile females	T	27	7.99***	.04	34.32****	.19	9.44***	.05	.89	.20	5.08*

* $p < .05$.
*** $p < .01$.
**** $p < .001$.

value scale reflect important concepts and ideas in the Jews' cultural and religious upbringing.

The final factor to be examined is status. In the comparisons involving race, belief, and status, status effects were significant in three of the eight samples. These three are all gentile samples (male Form A, $p < .001$; male, $p < .001$ and female, $p < .05$, Form T). In these samples, stimulus persons of high status are preferred to those of low status, although the proportion of variance contributed by the status effect is again minimal. In none of the four Negro samples was the status effect significant. No obvious explanation from Rokeach's theory is at hand as to why status should be more important for gentile than for Negro subjects. It may be that Negroes minimize the importance of status because of their limited opportunities to obtain high status.

The influence of status in analyses in which it is pitted against religious affiliation is no greater than in those in which it is pitted against race. (The rank-order difference between Ω^2 values for both status effects is not significant.) The status effect in the religious comparisons is significant in four out of eight samples. In the three of these four (all of which are Jewish samples) it is significant at $p \leq .001$ and in the other, at $p \leq .01$ (gentile females, Form T). Perhaps the relatively greater importance of status for Jewish samples may be understood in terms of the integral part that status attributes play in the value system held by upper-middle-class Jews, such as these subjects.

Seven of the 48 two-way interactions were significant. Four of these seven involve the Belief × Status interaction, and three of the four involve subjects who took Form T (see Table 9). A look at the mean responses to the various stimulus persons for these samples shows that when low status is ascribed to stimulus persons, similarity of values is essentially unrelated to subjects' responses. On the other hand, subjects tend to respond more favorably to a stimulus person of high status and like values than to one of high status and unlike values. The other three significant interactions are for Race × Belief (Negro females and gentile males, Form A) and Race × Status (gentile males, Form T). For both Race × Belief interactions, when unlike values are ascribed to stimulus persons, race is unrelated to subjects' responses. But when like values are ascribed to stimulus persons, the Negro sample preferred Negro to white stimulus persons and the gentile sample preferred white to Negro stimulus persons. In the Race × Status interaction, when lower status is attributed to stimulus persons race is unrelated to subjects' responses. When upper status, however, is ascribed to stimulus persons, subjects react more favorably to white than to Negro stimulus persons.

No interpretation is offered for these interactions, which appear to reflect complicated relationships between race, religion, and socioeconomic status of the subjects and the variety of meanings attached to potential associations with minority group members in a wide range of social situations. A separate analysis of the individual items on the social distance scale was undertaken to discover what some of these relationships might be.

Analysis of Individual Social Distance Items

Since subjects responded dichotomously (Yes or No) to the individual social distance items, the analysis of variance is inappropriate for these data. In order to present relevant comparisons for inspection, the percentage of endorsement for each level of each factor has been computed. The means and standard deviations for responses to the individual items of the social distance scale for the several stimulus persons, forms, and samples appear in Stein (1965, pp. 149–180). For a given item on the social distance scale, the total number of "Yes" responses to the four white stimulus persons was divided by the number of subjects responding to these stimulus persons, to obtain the absolute percentage of endorsement given to all white stimulus persons. This

process was repeated for responses to the four Negro stimulus persons, like-valued stimulus persons, etc. Looking at the first cell in Row one of Table 10, for example, we can say that for this sample of 25 Negro males, 84% of all responses to the four white stimulus persons were "Yes"; that is, subjects express themselves as quite willing to have a white stimulus person as a neighbor.

Tables 10 and 11 illustrate the absolute percentages for both Forms A and T. Since Tables 12 and 13 present the percentage difference between the two levels of each factor, these tables will be discussed in detail. The absolute percentage tables are presented in order to give an idea of the frequency with which items were endorsed. In general, the more intimate the social situation, the less frequently the item is endorsed.

Tables 12 and 13 show the percentage difference between the two levels of each of the three factors: race (or religion), belief, and status, on Forms A and T, respectively. These percentages were calculated by taking the difference between the two absolute percentages for the two levels of each factor (Tables 10 and 11). The percentage responding "Yes" to Negro stimulus persons was subtracted from the corresponding percentage for white stimulus persons, and also for unlike values versus like values, lower status versus upper status, and Jew versus Protestant or Catholic. Thus, the resultant percentage, if positive, reflects a preponderance of positive responses to white, like-valued, upper status, Protestant or Catholic stimulus individuals, respectively; and if negative, a corresponding preponderance of positive responses to Negro, unlike-valued, lower status, or Jewish stimulus individuals. For example, the percentage difference of −1 in the upper left cell in Table 12 means that 1% more positive responses were made to Negro than to white stimulus persons: the difference between the first two cells in Row 1 of Table 10 (84% and 85%). Sign tests were computed to test for differences between responses to the levels of a given factor and for sex differences. All values reported in this section are based on sign tests.

Table 12 presents the results for Form A for both racial and religious comparisons. A look at the columns headed Belief shows that the percentage differences are on the whole large and positive. In the comparisons involving Negro and white stimulus persons, belief appears to be as important with respect to items that identify casual situations as it does for items that identify more intimate ones.

With respect to differences associated with the race of the stimulus persons, among the two gentile samples, female subjects appear to be more tolerant racially than males. A sign test for this sex difference is significant at $p < .01$. For the gentile males, the percentage differences are moderately large and all positive; the males clearly tend to prefer white stimulus persons to Negroes ($p < .01$). In fact, none of the gentile male subjects would be willing to have a close relative marry any of the Negro stimulus persons (see Table 10). For the Negro samples, Table 12 tells a somewhat different story. Negro subjects prefer Negro stimulus persons to white for all items ($p < .01$), but the negative percentage differences are all relatively small. We also see in Table 12 that gentile subjects, particularly males, generally prefer high-status stimulus persons to low when status is pitted against race and belief. High is preferred to low status by both males and females ($p < .01$), but the difference is significant only for males. The two Negro samples show moderate preference for high- rather than low-status stimulus persons although the difference is not significant, and contrary to the results for the gentile samples, there is no significant sex difference.

The bottom half of Table 12 shows the results for the comparisons involving religion, belief, and status (Form A). With regard to religion, gentile males were more likely to reject Jews than were gentile females ($p < .01$). The percentage differences for religion are all quite small, but gentile females even show a net preference

TABLE 10

ABSOLUTE PERCENTAGES OF "YES" RESPONSES TO THE INDIVIDUAL ITEMS OF THE SOCIAL DISTANCE SCALE FOR EACH FACTOR LEVEL OF RACE (WHITE, NEGRO) OR RELIGION (PROTESTANT, CATHOLIC, JEWISH), BELIEF (LIKE VALUES, UNLIKE VALUES), AND STATUS (UPPER, LOWER), ADULT FORM

Item on social distance scale	Negro males $N = 25$						Negro females $N = 25$						Gentile males $N = 30$						Gentile females $N = 33$					
	W	N	LV	UV	US	LS	W	N	LV	UV	US	LS	W	N	LV	UV	US	LS	W	N	LV	UV	US	LS
Race × Belief × Status comparisons																								
Neighbor on street	84	85	100	68	88	80	72	77	93	57	71	79	88	62	78	72	83	67	76	65	91	50	71	69
Work on charity drive	67	70	88	49	75	62	56	60	78	38	60	56	67	63	78	52	72	58	66	76	91	51	66	76
Speaking acquaintance	76	79	85	71	83	72	65	79	86	58	72	72	73	62	78	57	77	58	74	76	84	65	71	78
Go to party	67	82	84	64	78	71	56	76	80	52	61	71	82	58	73	67	78	62	69	67	78	62	67	69
Member of social club	61	66	78	49	68	59	51	59	81	29	58	52	63	43	67	40	63	43	56	62	80	38	60	57
Live in same apartment house	80	83	86	77	83	80	73	77	87	63	78	72	78	58	67	70	72	65	80	66	85	62	75	71
Close personal friend	62	68	82	47	64	65	36	53	73	13	41	46	55	40	60	35	53	42	32	33	54	10	36	29
Invite home to dinner	66	80	84	62	74	72	52	70	82	40	63	58	68	40	63	45	67	42	42	34	60	17	40	36
Have close relative marry	33	49	47	35	43	39	25	33	44	14	29	29	47	0	35	12	28	18	30	12	33	9	27	15
Share an apartment with	36	43	51	28	37	43	23	42	52	12	29	36	33	13	32	15	28	18	27	22	45	4	27	23

Item on social distance scale	Jewish males $N = 69$						Jewish females $N = 68$						Gentile males $N = 32$						Gentile females $N = 30$					
	P	J	LV	UV	US	LS	P	J	LV	UV	US	LS	P/C	J	LV	UV	US	LS	P/C	J	LV	UV	US	LS
Religion × Belief × Status Comparisons																								
Neighbor on street	81	86	93	74	89	78	83	84	98	69	84	84	79	67	88	59	74	72	83	90	100	73	87	86
Work on charity drive	72	74	86	60	72	74	72	74	90	56	72	74	72	59	78	54	64	67	67	70	87	50	67	70
Speaking acquaintance	78	82	88	72	88	71	80	83	94	68	85	78	70	60	84	45	63	66	81	85	94	73	87	80
Go to party	82	82	88	76	93	71	82	90	96	76	87	85	83	84	91	76	80	87	78	87	90	74	87	78
Member of social club	57	70	78	50	70	57	66	71	90	47	71	66	66	56	81	42	63	60	61	62	80	44	63	61
Live in same apartment house	80	82	91	72	85	78	80	84	92	73	82	82	75	69	83	61	70	74	73	78	85	67	78	74
Close personal friend	46	54	74	25	56	43	39	41	71	10	42	38	53	38	76	15	46	45	43	36	62	16	45	34
Invite home to dinner	59	67	78	47	71	55	58	60	84	35	60	58	55	42	73	24	48	49	55	50	73	31	60	44
Have close relative marry	33	50	58	25	47	36	26	46	53	19	38	34	38	22	46	14	34	25	38	29	47	19	41	25
Share an apartment with	34	37	54	17	43	28	30	35	54	11	34	32	35	22	48	9	33	24	30	14	41	8	30	19

Note.—W = White J = Jewish
N = Negro P = Protestant
LV = Like values
UV = Unlike values
US = Upper status
LS = Lower status

P/C = Protestant/Catholic
Protestant subjects received
Protestant stimulus persons
Catholic subjects received
Catholic stimulus persons

TABLE 11

Absolute Percentages of "Yes" Responses to the Individual Items of the Social Distance Scale for Each Factor Level of Race (White, Negro) or Religion (Protestant, Catholic, Jewish), Belief (Like Values, Unlike Values), and Status (Upper, Lower), Teenage Form

Items on social distance scale	Negro males $N = 23$						Negro females $N = 24$						Gentile males $N = 26$						Gentile females $N = 26$					
	W	N	LV	UV	US	LS	W	N	LV	UV	US	LS	W	N	LV	UV	US	LS	W	N	LV	UV	US	LS
Race × Belief × Status comparisons																								
Sit next to in class	76	70	83	63	82	64	74	77	95	56	77	74	78	74	88	65	86	67	79	81	90	69	83	77
Work on committee with	68	48	77	39	66	50	60	60	83	38	69	51	55	69	76	48	77	47	63	65	81	48	75	54
Speaking acquaintance	56	50	68	39	58	48	62	68	78	52	68	62	65	74	82	56	75	64	69	79	86	62	81	67
Go to party to which person was invited	71	74	85	61	73	72	75	74	87	53	67	72	76	62	78	60	76	62	79	50	79	50	63	65
Eat lunch with	67	70	85	52	70	67	60	55	80	35	57	58	73	66	84	55	79	60	73	44	77	40	60	58
Member of social group	50	61	70	41	61	50	47	55	74	29	55	47	60	58	75	43	64	54	54	42	69	27	52	44
Live in same apartment house	59	53	66	46	59	52	79	68	87	59	78	68	64	46	64	45	59	51	75	38	58	56	60	54
Close personal friend	48	43	56	34	50	41	30	38	49	18	41	27	36	40	55	21	46	31	36	25	48	13	38	23
Invite home to dinner	43	46	59	30	50	39	44	53	70	27	55	41	44	29	48	25	48	25	40	21	44	17	36	25
Date brother (sister)	32	35	46	22	41	26	32	45	60	17	41	36	41	0	28	13	26	15	27	15	33	10	29	13

Items on social distance scale	Jewish males $N = 84$						Jewish females $N = 88$						Gentile males $N = 20$						Gentile females $N = 27$					
	P	J	LV	UV	US	LS	P	J	LV	UV	US	LS	P/C	J	LV	UV	US	LS	P/C	J	LV	UV	US	LS
Religion × Belief × Status comparisons																								
Sit next to in class	81	86	90	76	91	76	85	88	93	81	93	81	85	67	87	65	76	76	88	76	91	74	92	72
Work on committee with	65	71	82	55	83	53	66	64	83	48	77	53	54	54	64	43	60	48	68	58	80	46	74	52
Speaking acquaintance	82	82	92	73	85	79	85	86	95	76	90	81	74	76	83	67	87	63	74	71	85	59	78	67
Go to party to which person was invited	80	87	91	76	89	78	81	78	92	68	84	76	82	75	87	69	73	84	83	72	91	65	85	70
Eat lunch with	74	76	88	61	80	69	61	64	85	41	70	56	62	58	69	50	60	60	72	71	84	59	76	67
Member of social group	60	65	78	46	73	52	53	60	78	35	66	46	54	59	71	43	50	63	55	46	67	35	54	48
Live in same apartment house	77	84	88	75	85	77	82	87	89	80	87	82	74	69	78	65	74	69	81	61	74	68	72	71
Close personal friend	39	42	61	20	53	28	33	36	61	8	48	22	45	22	39	28	32	35	48	30	54	24	50	28
Invite home to dinner	45	49	66	28	57	37	46	47	72	22	57	37	40	30	38	32	32	37	55	36	65	26	52	39
Date brother (sister)	29	42	48	23	48	24	32	39	53	18	49	22	30	24	34	21	38	17	45	26	52	18	45	26

Note.—W = White J = Jewish
N = Negro P = Protestant
LV = Like values
UV = Unlike values
US = Upper status
LS = Lower status

P/C = Protestant/Catholic
Protestant subjects received
Protestant stimulus persons
Catholic subjects received
Catholic stimulus persons

TABLE 12

PERCENTAGE DIFFERENCES BETWEEN THE TWO LEVELS OF RACE OR RELIGION, BELIEF, AND STATUS FOR "YES" RESPONSES TO THE INDIVIDUAL ITEMS OF THE SOCIAL DISTANCE SCALE, ADULT FORM

Items on social distance scale	Negro males $N = 25$			Negro females $N = 25$			Gentile males $N = 30$			Gentile females $N = 33$		
	Race	Belief	Status	Race	Belief	Status	Race	Belief	Status	Race	Belief	Status
Race × Belief × Status comparisons												
Neighbor on street	−1	31	8	−5	37	−8	27	7	17	11	41	2
Work on charity drive	−2	39	13	−4	39	4	3	27	13	−9	40	−9
Speaking acquaintance	−3	14	10	−14	28	1	12	22	18	−2	19	−7
Go to party to which person was invited	−15	20	7	−20	28	10	23	7	17	2	14	−2
Member of social club	−4	29	9	−8	52	7	20	27	20	−6	42	3
Live in same apartment house	−3	10	3	−3	23	7	20	−3	7	14	23	4
Close personal friend	−5	35	−1	−15	60	−5	15	25	12	−1	44	7
Invite home to dinner	−13	22	2	−18	42	5	28	18	25	8	44	4
Have close relative marry	16	12	4	−8	30	0	47	23	10	18	24	12
Share an apartment with	−7	23	−6	−18	40	−7	20	17	10	5	40	4
	Jewish males $N = 29$			Jewish females $N = 68$			Gentile males $N = 32$			Gentile females $N = 30$		
	Religion	Belief	Status	Religion	Belief	Status	Religion	Belief	Status	Religion	Belief	Status
Religion × Belief × Status comparisons												
Neighbor on street	−4	19	11	−1	29	0	11	30	2	−7	27	0
Work on charity drive	−2	26	−3	−2	34	−2	13	24	−2	−3	37	−3
Speaking acquaintance	−3	16	17	−4	26	6	10	40	−3	−4	20	6
Go to party to which person was invited	0	12	22	−9	19	3	−1	15	−7	−9	16	9
Member of social club	−13	28	13	−4	43	4	10	39	3	−1	36	2
Live in same apartment house	−2	19	7	−4	19	0	6	23	−4	−5	18	5
Close personal friend	−8	49	13	−2	61	4	14	62	2	7	46	11
Invite home to dinner	−8	31	16	−2	49	2	12	49	−1	5	42	16
Have close relative marry	−17	32	10	−20	34	5	16	32	10	9	28	16
Share an apartment with	−3	36	14	−5	43	2	13	39	10	10	33	10

for Jews on the six least intimate items. One reason for this finding may be that Jewish students tended to predominate in the "leading crowd" in the Commutertown high school, as indicated by sociometric data in the study by Hardyck and Smith (in preparation). For Jewish subjects, the only item on which there is moderately strong in-group preference in both sexes is "have close relative marry." In general, status is not particularly important, but for the gentile samples it becomes more so for the more intimate items, as was not the case in the top half of the table. Jewish males also show moderately strong status effects on items throughout the scale, and these effects are significantly greater than are those for Jewish females ($p < .05$).

Corresponding percentage differences for the teenage form appear in Table 13. All the columns headed Belief again show large positive percentage differences, as none would expect from the results of the analysis of variance of total social distance scores. The findings with respect to Race for the gentile samples are different, however, from those with the adult form. Here, there is no significant sex difference in preference to white rather than Negro

stimulus teenagers. However, on one item: "Date your brother (sister)," gentile males show a much larger preference for the white stimulus persons as a date for their sibling than do the girls (41% versus 12%). This is clearly in line with societal demands and parallels findings on Form A for the item, "have close relative marry."

The present results are very similar to the findings of the Stein et al. (1965) study, which, it will be remembered Stein et al. involved gentile subjects but combined sexes for the analysis and used stimulus teenagers who were all high in status and varied only in race and belief. They found a significant belief effect on all 10 items (t tests could be employed in their design) and a significant race effect on only 3 items: "live in the same apartment house," "invite home to dinner," "date brother." The race effect also approached the .05 significance level on "go to a party with." They concluded that on such "sensitive" items race was important because the items reflected areas in which there are strong societal pressures against interracial contact. For the gentile samples in the present study, these same items show the largest differences in endorsements for white as versus Negro stimulus teenagers. The comparison of male and female subjects, new with the present study, shows that females are somewhat more likely than males to prefer white stimulus persons on the items concerning "live in the same apartment," "go to party with," and "invite home to dinner" but the reverse is true for "date your brother." In addition, female subjects show less acceptance of Negroes on the item, "eat lunch with." We can conclude, then, that the results of the Stein et al. study are essentially replicated by the present findings. For non-Jewish white subjects, belief is an important factor throughout the social distance scale, but race comes into play for the items that represent socially taboo areas of interracial contact.

For the Negro samples, similarity of belief is again quite important for interpersonal preference, with the female subjects tending to show slightly larger positive percentages for belief than the males ($p < .05$). Race effects on individual items are small and inconsistent.

The bottom half of Table 13 shows the percentage differences in response to teenage stimulus persons when religion, belief, and status are varied. Again, belief effects are all positive and fairly large for all items with all samples. The small religion effects may be summarized by saying that they are strongest for gentile females and for gentiles of both sexes for the more intimate items. The absence of appreciable religion effects on any of the items for the Jewish samples is perhaps surprising. For these samples, status is also moderately important, and becomes more so as the items increase in intimacy.

Responses to Stimulus Individuals Who Vary in Race or Religion and Status Factors Only

Stein et al. (1965) had found, in an analysis of responses of their subjects to a questionnaire that had been administered 2 months before the presentation of the stimulus persons varying in race and similarity of belief, that social distance to stimulus teenagers described in terms of race and status (with no information about belief) is determined by both of these factors, with the race effect explaining twice as much variance as the status effect.

Subjects in the present study had responded while in the eighth grade to similar stimulus teenagers or adults, with religion also varied. For teenage stimulus persons, status was varied as in the former study; for adults, it was varied by the combination of a professional occupation with the phrase "is making a good income" as opposed to a manual occupation with the phrase "is making a low income." For purposes of the present analysis, subjects were classified according to their own race and religion, and thus total social distance scale scores examined with

TABLE 13

PERCENTAGE DIFFERENCES BETWEEN THE TWO LEVELS OF RACE OR RELIGION, BELIEF, AND STATUS FOR "YES" RESPONSES TO THE INDIVIDUAL ITEMS OF THE SOCIAL DISTANCE SCALE, TEENAGE FORM

Items on social distance scale	Negro males $N = 23$			Negro females $N = 24$			Gentile males $N = 26$			Gentile females $N = 26$		
	Race	Belief	Status	Race	Belief	Status	Race	Belief	Status	Race	Belief	Status
Race × Belief × Status comparisons												
Sit next to in class	6	20	19	−2	40	2	4	23	20	−2	21	6
Work on committee with	19	38	16	0	46	17	−14	28	30	−2	33	21
Speaking acquaintance	6	29	10	−7	26	5	−8	26	11	−10	25	13
Go to party to which person was invited	−2	24	2	−9	34	−5	14	17	14	29	29	−2
Eat lunch with	−3	33	2	5	45	−1	7	30	20	29	36	2
Member of social group	−12	28	11	−8	45	8	1	32	10	12	42	8
Live in same apartment house	6	20	6	11	28	10	19	19	7	36	2	6
Close personal friend	4	22	8	−8	31	14	−5	34	15	12	35	15
Invite home to dinner	−3	29	11	−10	43	14	14	22	23	19	27	12
Date brother (sister)	−2	24	15	−13	42	5	41	14	10	12	23	15
	Jewish males $N = 84$			Jewish females $N = 88$			Gentile males $N = 20$			Gentile females $N = 27$		
	Religion	Belief	Status	Religion	Belief	Status	Religion	Belief	Status	Religion	Belief	Status
Religion × Belief × Status comparisons												
Sit next to in class	−5	14	15	−3	12	12	18	22	0	12	16	20
Work on committee with	−6	27	30	2	35	24	0	22	12	11	33	22
Speaking acquaintance	0	19	6	0	19	8	−2	16	24	3	26	11
Go to party to which person was invited	−7	15	11	3	24	8	7	18	−11	11	26	15
Eat lunch with	−2	27	11	−3	44	14	4	19	0	1	24	9
Member of social group	−5	32	21	−6	43	20	−5	28	−12	9	32	6
Live in same apartment house	−7	12	8	−5	8	5	6	13	6	20	6	2
Close personal friend	−2	40	25	−3	52	26	22	11	−3	18	30	23
Invite home to dinner	−5	38	20	0	50	20	10	6	−5	20	39	14
Date brother (sister)	−13	25	24	−7	36	28	6	13	21	18	34	19

respect to the stimulus persons specified in Table 14.[5]

This analysis will be discussed in terms of a summary of the 24 analyses of variance computed on these data (Table 15). Means and standard deviations are presented in Stein (1965, p. 80).

Looking first at the top half of Table 15, we can see that the race effect was significant at the .001 level in 9 of the 12 samples, and at lower levels in 2 other samples (Negro females, Form T, $p < .01$; Negro males, Form A, $p < .05$). It fell short of significance in only one sample (Negro male, Form T). Examination of the means shows that Negro subjects tend to prefer stimulus persons of their own race. Inspection of the column in Table 15 showing the proportion of variance contributed, however, reveals that race seems to be more important for white samples than for Negro samples. (N is too small to compute a sign test.)

When pitted against race, status has

[5] For Catholic subjects, race and religion were confounded in the descriptions of the Negro stimulus persons. Since there was no "Catholic Negro," the subjects had to respond to a "Protestant Negro." Only 591 of the 630 subjects had scores on the appropriate social distance scales in the interest and attitude questionnaire.

significant effects in only 7 of the 12 pertinent samples, and in none of these does significance approach the .001 level. The rank-order difference between the Ω^2 values for the race and status effects shows that race contributes significantly more variance than status ($p < .01$). In comparison to the race effect, then, status is a less powerful but still important determinant of choice. These results confirm the findings of Stein et al.

Status is strongly affected by the form of the questionnaire. Only two of the six analyses on the adult form as compared with five of six analyses on the teenage form yield significant results. The rank-order difference between Ω^2 values for Form T versus Form A is significant at less than the .05 level; more variance is contributed by status in Form T samples than in Form A samples. There seem to be two possible explanations for these findings. First, the status descriptions in the adult form may have been too vague ("making a good income"; "making a low income"). The other possibility is that adult status attributes are less important to teenagers than teenage status attributes. We may conclude, then, that in the absence of information about beliefs, race is a powerful determinant of interpersonal preference with status contributing a smaller but significant influence that is confined primarily to the teenage form. In general, there are few significant interaction effects.

The bottom half of Table 15 presents the results of analyses in which religion and status are varied. All samples show highly significant religion effects (nine reach the .001 level and the other three the .01 level). The three samples in which the religion effect is only significant at $p < .01$ are all Protestant samples. This finding would follow if religious affiliation is somewhat more salient for the "minority" Jewish and Catholic subjects than for the "majority" Protestants. In 10 of the 12 samples there are significant status effects, with 6 of these being significant at $p < .001$. The rank-order difference between the Ω^2 values for religion and status was not significant. Status

TABLE 14

Stimulus Persons Used in Analyses in Which Race or Religion and Status Are Varied

Race or Religion of Subject	Race Varied	Religion Varied
	Stimulus Persons Used in Analysis	Stimulus Persons Used in Analysis
Protestant	White Protestant Negro Protestant	White Protestant White Jewish
Catholic	White Catholic Negro Protestant	White Catholic White Jewish
Jewish		White Protestant White Jewish
Negro	White Protestant Negro Protestant	

differences thus appear to be more important when status is varied with religion than when it is varied with race—a finding that did not appear in the analyses reported earlier in which similarity of belief was also varied. This finding is not too surprising, in that religion is a less powerful variable than race, and, of course, belief. When only religion and status are varied, subjects are as likely to make distinctions on the basis of one factor as they are on the other. When race and status are varied, the factor of race predominates, and, as we have seen, when information about belief is added, this tends to wash out the influence of the other factors.

Again, the status factor seems particularly important for all four Jewish samples, which show status effects significant at $p < .001$ and have status contributing a large proportion of the variance. Only 4 of the 36 interactions are significant.

Implications

In a full-scale test of Rokeach's theory of belief prejudice with ninth-grade students, the present results point overwhelmingly to the validity of the theory. When information about a stimulus person's beliefs in the area of personal values is made available, *perceived* similarity—or dissimilarity—in beliefs is the

TABLE 15

SUMMARY OF THE 24 ANALYSES OF VARIANCE FOR RESPONSES TO THE TOTAL SOCIAL DISTANCE SCALES OF THE INTEREST AND ATTITUDE QUESTIONNAIRE WHICH WAS GIVEN WHEN THE STUDENTS WERE IN THE 8TH GRADE

Sample	Form	*N*	Race		Status		Individual × Race	Individual × Status	Race × Status
			F	Prop. of variance	*F*	Prop. of variance	*F*	*F*	*F*
Negro males	A	23	5.31*	.03	1.41	.00	1.55	2.31	.22
Negro females	A	18	18.51****	.17	.33	.00	3.66	1.26	.41
Protestant males	A	23	19.01****	.16	8.05***	.02	6.70*	2.35	1.15
Protestant females	A	26	40.41****	.28	4.96*	.00	5.54*	.82	1.08
Catholic males	A	27	29.87****	.16	1.54	.00	3.79	2.38	4.26*
Catholic females	A	30	29.18****	.22	1.41	.00	5.97*	1.46	.51
Negro males	T	18	3.26	.01	6.43*	.08	1.43	3.44	.15
Negro females	T	20	11.32***	.09	5.64*	.03	3.15	2.28	1.43
Protestant males	T	15	26.17****	.16	13.89***	.14	1.40	2.48	.24
Protestant females	T	16	24.35****	.36	11.89***	.03	8.66*	1.67	0.0
Catholic males	T	28	26.37****	.14	7.17*	.03	2.09	1.82	0 0
Catholic females	T	31	69.07****	.31	2.30	.00	3.44	2.99	.01
			Religion		Status		Individual × Religion	Individual × Status	Religion × Status
Jewish males	A	52	39.26****	.11	33.64****	.09	2.42	2.53	0.0
Jewish females	A	55	62.09****	.20	16.34****	.04	3.66	2.66	.07
Protestant males	A	23	12.04***	.08	4.54*	.01	3.42	1.86	.09
Protestant females	A	27	14.89****	.08	2.85	.01	2.78	2.85	.22
Catholic males	A	27	28.99****	.18	7.54***	.02	2.06	.81	.17
Catholic females	A	32	22.26****	.22	2.38	.00	24.39****	1.82	4.35*
Jewish males	T	74	37.71****	.04	102.90****	.30	2.54	6.95*	3.50
Jewish females	T	73	28.36****	.04	69.76****	.22	2.37	5.34*	.43
Protestant males	T	15	16.58***	.07	16.14***	.21	.83	2.64	2.46
Protestant females	T	16	8.78***	.10	16.87****	.14	3.82	2.47	3.55
Catholic males	T	27	39.96****	.23	15.78****	.08	1.64	1.45	.97
Catholic females	T	32	22.12****	.12	6.73*	.03	4.05	4.14	2.13

* $p<.05$.
*** $p<.01$.
**** $p<.001$.

primary determinant of attitudes of white gentiles toward Negroes and Jews. Likewise, knowledge of belief systems, when it is made available, is the most important factor in Negro and Jewish students' attitudes toward members of the majority. Only secondarily does racial or religious membership per se, or high versus low relative socioeconomic status, influence the students' feelings and action or orientations toward others under these circumstances.

The generality of the findings is impressive. These results hold up for both teenage and adult forms of the questionnaire, for both sexes, and for Negro, Catholic, Protestant, and Jewish subjects as well as for ninth-grade students in California (Stein et al., 1965) questioned about Negro stimulus persons, in the absence of any substantial opportunity for interaction with Negroes.

It is important to elaborate on these findings since they initially give the appearance of opposing common notions of prejudice. In conventional accounts, so much emphasis has been placed on ethnic membership per se as a determinant of

prejudice toward members of minority groups that at first it seems incredible that belief incongruence could be the major determinant of prejudice. Yet these rather striking results can be reconciled with the practical importance of ethnic membership in social life.

In the first place, the present study may confront gentile teenagers for the first time with information that tells them that there are Negroes who believe in many of the same things that they themselves do; that these Negroes hold many values that they themselves consider of vital importance. We are asking students to make decisions about their feelings towards and willingness to interact with Negroes whom they have not before evaluated from this point of view. In a sense, it is like asking them, "If Negroes were more like you than you think they are and believed in the same things you do, would you then like them?" Our data give an affirmative answer to this question, particularly in regard to "feelings" toward Negroes, but the answer must be qualified by a second important feature of the data.

The students do make a distinction between situations that are relatively free from strong societal pressures, on the one hand, and ones that represent areas of interracial contact in which societal taboos are continually reinforced, on the other. From the data for the individual items of the social distance scale, it appears that gentile subjects are willing to interact with like-valued Negroes in such situations as "sit next to in class," "work on a committee with," "have as a speaking acquaintance," "eat lunch with," and even "have as a member of one's social group" or "have as a close personal friend." However, in situations of culturally defined intimacy or in which parents or other adults would be visibly involved in the contact, the subjects show great reluctance to interact with Negroes. Thus, on such items as "invite home to dinner," "live in the same apartment house," "date brother," "have a close relative marry," and even "have as a neighbor on the street," gentile subjects are much more prone to react in racial terms, frequently rejecting contact with Negroes.

This finding appears to give partial support to Triandis' (1961) criticism of Rokeach's theory. Triandis claimed that we object to having a Negro live next door to us because he is Negro, not because of what he believes. Since this type of situation is one in which societal pressures discourage interracial contact, belief is a less powerful determinant than race for some subjects.

Rokeach[6] suggests, however, that we still need not invoke an interpretation based on racial criteria for these kinds of situations. Instead, he claims that the principle of belief congruence is applicable but not in terms of what the Negro is seen to believe. In such a situation as having a Negro live next door, Rokeach suggests that the white person believes that the presence of Negroes in the neighborhood would affect the rise and fall of real estate values. This belief, therefore, could account for not desiring a Negro as a next door neighbor but could be quite independent of the white person's attitude toward the Negro. Rokeach offers an additional example: "Suppose a white person refused to sit down next to a Negro on a bus in Montgomery, Alabama. Is it because that person is black, because of certain beliefs that he sees the Negro to have, or because the white person *believes* that if he sits down the bus driver will ask him to get off the bus, or *believes* that he will be breaking a law for which he can be arrested?" The question may well be raised, however, whether a belief theory of prejudice thus extended is capable of empirical disconfirmation. The present study has found strong support for the more restricted version of the theory as the major though not exclusive determinant of interpersonal prejudice.

Negro subjects, in responding to white stimulus persons, make few if any of the distinctions that whites do in the situations described on the social distance scale. In general, they would seem to have little to

---

[6] Personal communication.

lose and much to gain from social relations with whites.

The results for the religious comparisons do not follow this general interpretation. Although there are some subjects who balk at close personal contact with stimulus persons of another religion, in almost all cases difference in belief is the crucial factor in determining interreligious relations. There are fewer "taboo areas" for interreligious contact than for interracial contact. In the case of the former, only the items, "invite home to dinner," and "have a close relative marry," and possibly, "date brother," seem to reflect socially strained areas of contact.

The interpretation of the Rokeach and Triandis controversy offered by Stein et al. (1965) thus seems to be strengthened by the present findings. Knowledge of belief systems, if they reflect belief congruence, leads many more gentile subjects to evaluate their feelings and potential behaviors toward Negroes in a positive manner. Without this knowledge, Negroes are assumed to have dissimilar beliefs and values and are consequently rejected. Even when information about belief systems is supplied, there are still some subjects who either feel bound by societal pressures or genuinely harbor hostile feelings toward Negroes and refuse to interact with them particularly in areas of intimate contact. It is not surprising, then, that when information about beliefs is absent and only race and status are varied, racial considerations become dominant.

Some cautionary remarks are in order, in view of the compelling consistency of the findings. This study ignores important conditions in social reality which might well mitigate against our findings. The theory of belief prejudice needs to be tested in conjunction with variations in the social forces which heavily influence the formation and maintenance of prejudiced attitudes (see Rokeach & Mezei, 1966, for a beginning in this direction). Moreover, the theory needs cross-validation in both southern and northern communities wherein racial strife is a common occurrence and where *opportunities* to perceive similarity of beliefs can be limited by conditions in the social structure.

While the values of the factors of race and religious affiliation in this study are absolute (e.g. white, Negro, etc.), the value items representing the belief factor, and the status attributes, are arbitrarily set. Other possible values for these factors need to be sampled before we can know the limits to which the present findings can be generalized.

Caution is also required in regard to possible effects of the relative salience with which information about race, religion, and belief was presented in the experimental materials. Race or religion was indicated by a mere word whereas the information on belief required an entire page. One could also, certainly, have played down the importance of belief by using less relevant values or by reducing the amount of contrast between similar and dissimilar values. For that matter, the salience of race could have been increased by adding pictures of the stimulus persons. In addition, for these northern subjects, it is certainly "socially undesirable" for them to stress race per se especially considering the intellectualism of the method and setting. Moreover, Triandis and Davis (1965) were able to obtain powerful race effects with subjects classified independently as "racially prejudiced" and with the use of social distance scale items reflecting negative behaviors such as "exclude from the neighborhood." In a sense, then, the results are specific to the method used and need further checking to see how much they are tied to the method. To show that one can pick evaluative beliefs, however, that so predominate over race is essentially to support Rokeach's theory, even though there are other beliefs for which the prediction might not hold. Knowledge of the perceived similarity of belief systems is clearly a crucial factor in the understanding of prejudice. Many possible strategies for the solution of racial and ethnic tensions follow from this fact.

REFERENCES

Biomedical computer programs. *The general linear hypothesis.* (BMD 14), University of California, Division of Biostatistics, Los Angeles, 1961.

Byrne, D., & Wong, T. J. Racial prejudice, interpersonal attraction, and assumed dissimilarity of attitudes. *Journal of Abnormal and Social Psychology*, 1962, **65**, 246–253.

Edwards, A. L. *Statistical methods for the behavioral sciences.* New York: Rinehart, 1954.

Hays, W. L. *Statistics for psychologists.* New York: Holt, 1963.

Morris, C. *Varieties of human value.* Chicago: University of Chicago Press, 1956.

Rokeach, M. Belief versus race as determinants of social distance: Comment on Triandis' paper. *Journal of Abnormal and Social Psychology*, 1961, **62**, 187–188.

Rokeach, M., & Mezei, L. Race and shared belief as factors in social choice. *Science*, 1966, **151**, 167–172.

Rokeach, M., Smith, P. W., & Evans, R. I. Two kinds of prejudice or one? In M. Rokeach (Ed.), *The open and closed mind.* New York: Basic Books, 1960. Pp. 132–168.

Stein, D. D. Similarity of belief systems and interpersonal preference: A test of Rokeach's theory of prejudice. Unpublished doctoral dissertation, University of California, Berkeley, 1965.

Stein, D. D., Hardyck, J. A., & Smith, M. B. Race *and* belief: An open and shut case. *Journal of Personality and Social Psychology*, 1965, **1**, 281–289.

Triandis, H. C. A note on Rokeach's theory of prejudice. *Journal of Abnormal and Social Psychology*, 1961, **62**, 184–186.

Triandis, H. C., & Davis, E. E. Race and belief as determinants of behavioral intentions. *Journal of Personality and Social Psychology*, 1965, **2**, 715–725.

Winer, B. J. *Statistical principles in experimental design.* New York: McGraw-Hill, 1962.

(Received August 24, 1965)

Vol. 80, No. 9 Whole No. 617, 1966

Psychological Monographs: General and Applied

BODY ATTENTION PATTERNS AND PERSONALITY DEFENSES[1]

SEYMOUR FISHER

State University of New York, Upstate Medical Center, Syracuse

Multiple studies were pursued of the hypothesis that the manner in which an individual distributes attention to his body is linked with his traits and personality attributes. Attitudes involving the following body awareness dimensions were measured: right versus left, front versus back, eyes, total body, and heart. A variety of personality parameters were sampled by means of questionnaires, selective memory responses, and semiprojective reactions to pictures. Conflictual feelings about certain wishes and aims were also evaluated from responses to stimuli presented in the Ames Thereness-Thatness apparatus. College students and psychiatric patients constituted the samples studied. It was possible to demonstrate that heightened awareness of specific body sectors is accompanied by characteristic conflicts and modes of defense.

It has been proposed that an individual's body experiences reflect the nature of his personality defenses. Freud's (1924, 1938) descriptions of the oral and anal character types depict an explicit relationship between investment of energy in certain body sectors and the existence of specific conflicts and defenses. Related equations between body feelings and personality patterns have been proposed by Schilder (1935), Reich (1949), Alexander (1948), and others (Ferenczi, 1916; Fenichel, 1945). In almost all of these systems it is assumed that personality variables are correlated with attitudes toward particular body sectors as a function of one or both of the following considerations:

1. It is suggested that in the course of socialization the child acquires certain response patterns (e.g., traits) because of crucial experiences he has with his parents. These experiences often revolve about body functions linked with specific areas of his body and result in his placing special valuations upon these areas. Thus, his style of sexual behavior might be influenced by the orientation he adopts from his parents toward the sexual regions of his body (e.g., lower half of body). Consequently, there would be a correlation between his attitudes toward sexual expression and his attitudes toward the sectors of his body having sexual functions.

2. Another consideration which has been noted is the unique closeness of the individual's body to himself as a perceiver. His body is the only object in his perceptual field which he simultaneously perceives and is also a part of himself. Its special closeness to himself (ego, identity) maximizes the likelihood that it will reflect and share in his most important preoccupations. Like all ego significant objects it can become a convenient "screen" upon which are projected one's most salient concerns. An example of such projection would be provided in the case of the individual who feels unimportant and inferior and then presumably transfers this view to his body by perceiving it as smaller than it is. Indeed, Popper (1957) and also Wapner and Krus (1959) have shown that failure experiences result in subjects perceiving themselves as relatively shorter in stature.

The discovery of relationships between personality variables and body attitudes would open many possibilities. It would help to clarify the role of body experience in personality processes. Also, it would permit a new approach to evaluating personality variables based not on their direct measurement but rather in terms of their body attitude representations. Information about relationships between body

[1] This study was partially supported by United States Public Health Grant M-5178 and also National Science Foundation Grant GP-1137.

attitudes and personality variables is only beginning to become discernible. By way of resumé, the following studies may be mentioned. One series of projects has established that an individual's mode of experiencing the boundary regions of his body (viz., skin and muscle) is linked with traits relating to self-assertion, self-expression, and mastery (Fisher, 1963; Fisher & Cleveland, 1958). Some tentative findings are available concerning the size one ascribes to one's body and the degree to which one is aggressive (Fisher, 1964b) and also field dependent (Epstein, 1957). Dissatisfaction with one's body has been shown to be related to feelings of insecurity (Jourard & Secord, 1955). There are also reports which variously: tie in attitudes toward the right and left sides of one's body with one's degree of hostility toward the opposite sex (Fisher, 1965a); relate body awareness to narcissism (Secord, 1953); and demonstrate a relationship between "strength of body image" and "effective control of the primary process [Reiff, 1962]." These studies range widely and revolve about diverse measuring procedures. With few exceptions they deal not with attitudes toward specific regions of the body but rather with broad, abstract body-image dimensions.

The present project was concerned with pursuing the personality correlates of a series of body-image parameters based on a common rationale. Of central interest was the question whether differences in the relative distribution of an individual's attention to various parts of his body or to his body as contrasted to nonbody objects are accompanied by corresponding personality differences. Do body-attention patterns provide meaningful information about the personality structure? Several different approaches to this issue were undertaken.

RIGHT-LEFT

The first approach was concerned with the distribution of attention to the right versus left body sides. It was anticipated that the relative prominence of the right and left sides in an individual's body scheme (as defined by focus of attention) would be correlated with indexes of sexual adjustment and sexual identification. There were several reasons for anticipating such correlates. One finds a long history of anecdotal and clinical observation (Fisher, 1965a) suggesting that the right-left dimension relates to matters of masculinity and femininity. It has been assumed that the right is symbolic of masculinity and strength and the left of femininity and weakness. While the few studies dealing with the right-left distinction have not supported this particular formulation, they indicate that right-left is pertinent to sex role issues. Fisher (1965a) reported that differences in number of male and female names applied to puppets placed upon the right and left hands were related to attitudes regarding the relative superiority of men to women. Similarly, the ability of subjects to make a clear distinction in the apparent sizes of their right and left hands while wearing aniseikonic lenses has proven to be correlated with projective indexes tapping sex role variables (Fisher, 1960b). It is of interest too that boys perceive autokinetic movement as more right-directional than do girls (Fisher, 1962).

Study 1A

In first approaching the possible correlates of the distribution of attention to the right and left body sides, an exploratory study was undertaken which involved relating a right-left attention index to a gross measure of the individual's sexual orientation, that is, his level of heterosexual interest. Two questions were under consideration: (*a*) Is the degree of right versus left attention correlated with heterosexual interest? (*b*) If so, what variations in such interest accompany greater focus on the right as compared to the left sides?

Procedure. Relative direction of attention to the right and left body sides was evaluated by means of the Body Focus Questionnaire. This is an instrument (Fisher, 1964b) which involves the subject comparing his degree of awareness of a series of different sites on his body. It presents him with a list of 91 paired references to body sectors (e.g., right hand versus left hand, left leg versus right leg, stomach versus arm, arm versus neck). He is asked to turn his attention upon his body and to

indicate for each pair of body parts which he is "most conscious of or aware of right now." Nine of the comparisons involve a right-side sector versus a left-side sector. The remaining items can be scored for other body dimensions, but in the present study served as filler to conceal the fact that the measurement process was concerned with the right-left dimension. A subject's relative degree of focus upon the right side of his body could range from 0 through 9. Administration of the Body Focus Questionnaire took place in a group setting.

The Edwards Personal Preference Schedule was administered to obtain responses to the Heterosexuality scale which inquires concerning the degree to which the subject considers heterosexual activities (e.g., to be in love, to kiss the opposite sex) as characteristic of himself.

Subjects. The subjects were 51 male college students who were paid a fee for participating. Their median age was 20. They were all right-handed in order to eliminate the possible effects of handedness upon the distribution of right-left attention. The study was restricted to males because data from other sources already cited (Fisher, 1962) clearly suggest that sex differences are to be expected with regard to the significance of right-left.

Results. The mean Body Focus Questionnaire Right score was 5.5 ($\sigma = 2.3$). The mean Edwards Heterosexuality percentile score was 53.6 ($\sigma = 30.1$). There proved to be a product-moment correlation of $-.27$ ($p = .05$) between the Right score and the Heterosexuality score. Thus, the greater the attention an individual focuses on the right side of his body the less is his apparent heterosexual orientation as defined by the Heterosexual scale.

Discussion. One sees initial support for the view that the right-left dimension of the body image reflects aspects of the individual's sex role and sexual adjustment. It is an intriguing question as to why focus of attention on the right rather than the left side should denote a reduced level of heterosexual response. Discussion of this issue will be postponed until a later point.

Study 1B

An attempt was made to generalize from the initial findings by formulating the following hypotheses which specify disturbance in various levels of one's sexual behavior as correlated with increasing emphasis upon the right side of the body:

The greater the focus of a man's attention upon the right side of his body:

1. The less active will be his general level of heterosexual behavior.

2. The more defensive he will be when confronted with stimuli that arouse anxiety about sexual identity.

3. The more anxious he will be in responding to symbolic representations of sexual problems and threats.

Procedure. The Body Focus Questionnaire (BFQ) was used to determine the subject's right-left distribution of attention. To increase the reliability of the measure it was administered on two separate occasions, with an average of 7 days intervening. A total Right score was derived equal to the sum of the right-side sites chosen on the two different occasions. The product-moment correlation between the test and retest Right scores was .58

Multiple procedures were employed to get at the sexual role variables referred to in the hypotheses.

Heterosexual activity level was appraised directly by means of a questionnaire inquiring concerning the subject's present and past sexual interaction with girls. He was asked to indicate the age at which he began dating; the average number of dates he had during each of the 4 years of high school and also currently; and the number of times he had "gone steady" or been engaged. The following scores were derived from this information:

1. Age at which began dating.

2. Average of number of dates per week for the 4 years of high school.

3. Average number of dates per week in current life.

4. A "serious dating" index equal to the number of times has "gone steady" (maximum of 2) plus a credit of 1 for currently going steady plus a credit of 2 for being currently engaged.

Another procedure was used to determine the subject's reactions to stimuli intended to arouse anxiety about sexual identity. The stimuli were 15 pictures of human figures from a series developed by Doidge and Holtzman (1960) to study homosexuals. The figures were taken from drawings, paintings, and statues photographically reproduced so as to make the sex ambiguous. The 15 selected for the present study were maximally ambiguous in this respect. It was presumed that the greater an individual's uncertainty about his sexual identity the more disturbing he would find such pictures and therefore the more defensive negative affect they would arouse in him (e.g., as suggested by Murray [1938]). With this rationale, a procedure was devised to involve the subject judgmentally with each of the pictured figures and to record his reactions via ratings.

Each picture was projected in a semidarkened room for 15 seconds. The subject was asked to de-

TABLE 1

MEANS AND STANDARD DEVIATIONS FOR BFQ RIGHT SCORES, INDEXES OF HETEROSEXUAL BEHAVIOR, VAGUE SEX PICTURE RATINGS, BLACKY PICTURE RANKS, AND SEXUAL REFERENCE SCORES

Variable	Mean	σ	N
BFQ right	10.5	4.1	49[a]
Heterosexual behavior			
Age began dating	14.7	1.7	50
Dates in high school	.8	.5	52
Current dates	1.7	1.3	49
Serious dating score	2.3	1.7	51
Vague sex pictures			
Ugly ratings	44.0	5.5	48
Unfriendly ratings	43.7	4.7	48
Unfriendly plus ugly	87.9	8.5	48
Blacky pictures			
Oedipal intensity	5.1	2.6	51
Castration anxiety	7.0	2.7	51
Love object	4.2	2.8	51
Sexual references	1.8	1.3	49

[a] Variations in N are a function of subjects either missing certain tests or not answering specific questions in a given test.

cide whether the figure was male or female and to indicate on a 5-point scale its apparent degree of masculinity-femininity. These judgments were obtained to insure there would be direct confrontation with the threatening, poorly defined sexual attributes of each figure. To measure the amount of defensive negative affect aroused, two other ratings of each were obtained. One was an evaluation of the attractiveness of the figures on a 5-point scale ranging from "ugly" to "good looking" and the other was a rating of the friendliness of the figures on a 5-point scale ranging from "friendly" to "unfriendly." From the ratings three indexes of defensive negative response was computed:

TABLE 2

PRODUCT-MOMENT CORRELATIONS OF BFQ RIGHT SCORES WITH HETEROSEXUAL ACTIVITY INDEXES

BFQ Right versus	r	Significance level
Age began dating	.30 (N = 50)	<.05
Average number dates per week in high school	−.21 (N = 52)	>.10
Average number of dates per week currently	−.39 (N = 49)	<.01
Score for serious dating	−.29 (N = 51)	<.05

1. Sum of the 15 ratings of degree of ugliness.
2. Sum of the 15 ratings of degree of unfriendliness.
3. A total score equal to the sum of the ugliness and unfriendliness ratings.

A third technique was devised to tap anxiety about sexual issues and conflicts in terms of responses to the Blacky Pictures (Blum, 1949), which consist of 11 scenes in which a dog is portrayed as engaged in activities illustrating crucial developmental problems in the psychoanalytic scheme (e.g., Oral Eroticism, Castration Anxiety). A new approach to measuring response to the pictures was attempted. The subject is presented with the 11 pictures arranged in the order in which they are usually administered. He is told that in a later session he will be asked to compose stories about some of them. In preparation for this, he is to examine the pictures and put them in rank order, with the one he would most prefer to elaborate upon first in sequence and the one he would least prefer to describe last in sequence. It was assumed the greater the anxiety aroused in a subject by a picture the more motivated he would be to evade involvement with it by assigning it a low rank order. Only the reactions to three pictures specifically related to heterosexual issues were considered pertinent for the present study. The three pictures were as follows: Oedipal Intensity, Castration Anxiety, and Love Object. It was anticipated that the greater the attention devoted to the right side of the body the lower would be the rank order assigned to these pictures.

Another procedure was used to determine how easily the individual verbalizes sexual thoughts in a free expressive situation. It was presumed that the greater his anxiety about sexual matters the more difficult it would be for him to make sexual references. The opportunity for free expression of thoughts and images was provided by asking him to list on a sheet of paper "20 things you are conscious of or aware of right now." Responses were obtained on two occasions, with a week intervening. Sexual references were defined to include only direct statements about heterosexual interests or activities (e.g., "I would like to kiss a girl," "I am going on a date tonight"). Two judges achieved 91% agreement in their scoring of 40 protocols. The Sex Reference score equaled the sum of sexual references in the two samples of responses and could range from 0 to 40. It was anticipated that BFQ Right would be negatively related to the Sex Reference score.

A new sample of subjects was used consisting of 49 male college students with a median age of 20. The BFQ Right score was determined by means of a revised scale with 15 items instead of 9.

Results. The BFQ Right score proved to be correlated in the predicted direction with the subjects' reports of heterosexually motivated behavior. As shown in Table 2,

TABLE 3
Product-Moment Correlations of BFQ Right Scores with Indicators of Negative Response to Vague Sex Pictures

BFQ Right versus	r	Significance level
Sum of ugly ratings	.30 (N = 48)	<.05
Sum of unfriendly ratings	.30 (N = 48)	<.05
Unfriendly and ugly	.36 (N = 48)	.01

the higher his BFQ score the later the age at which he began to date girls ($r = .30$, $p < .05$); the less his current amount of dating per week ($r = -.39$, $p < .01$); and the lower his score for serious dating as defined by "going steady" and being engaged ($r = -.29$, $p < .05$). There was also a correlation of $-.21$ between BFQ Right and amount of dating in high school which is in the predicted direction, but not significant ($p > .10$).

When one examines the relationships between BFQ Right and the Vague Sex Picture rating, it is apparent that they are supportive of the hypothesis under consideration. Table 3 indicates that each of the Vague Sex Picture ratings (Ugly and Unfriendly) is correlated .30 with BFQ Right ($p < .05$) and that the combined Ugly and Unfriendly ratings attain a correlation of .36 ($p = .01$) with BFQ Right.

TABLE 4
Product-Moment Correlations of BFQ Right Scores with Blacky Picture Rank Scores and Sex Reference Scores

BFQ Right versus	r	Significance level
Blacky		
Oedipal intensity	.05 (N = 51)	n.s.
Castration anxiety	.22 (N = 51)	>.10
Love object	−.13 (N = 51)	n.s.
Sex reference		
Set 1	−.30 (N = 50)	<.05
Set 2	−.38 (N = 50)	<.01
Sum	−.39 (N = 50)	<.01

The results shown in Table 4 indicate that BFQ Right has a chance relationship with the Blacky Pictures. Only the correlation between BFQ Right and Castration Anxiety approaches significance in the predicted direction ($r = .22$, $p > .10$).

Correlations between BFQ Right and the Sex Reference scores were significant in the predicted direction. BFQ Right had a correlation of $-.30$ ($p < .05$) with Sex References in the first set of responses; $-.38$ ($p < .01$) with Sex References in the second set of responses; and $-.39$ ($p < .01$) with the sum of Sex References in both sets.

Discussion. The findings support the proposition that the greater a man's focus of attention upon the right as contrasted to the left side of his body the more likely he is to be characterized by inhibition in his heterosexual behavior, anxiety about his sexual role, and difficulty in expressing ideas with sexual reference. It is true that the Blacky Pictures data were of a chance order and therefore not congruent with the original hypotheses. Apparently, the sex-role difficulties revealed in most of the procedures employed are not detected by the Blacky Pictures.

A prime question raised by the findings is why heterosexual difficulties are associated with a focus of attention on the right as opposed to the left body side. Two possible explanations will be offered. One derives from observations concerning the differential response characteristics of the right and left sides. The response of the right side tends to be slower and more controlled than that of the left. Schoen and Scofield (1935) reported that when the eyes of the right-eyed person are shifting from one fixation point to a new target, the left eye responds first, "snapping" to its new position and sometimes overshooting, as compared to the right eye, which

moves more gradually and precisely. Similarly, Travis and Herren (1929) and Jasper (1937) noted that when right-handed individuals were asked to perform a task quickly and simultaneously with both arms, the left responded first. Such findings suggest that in right-handed persons the right side is characterized by a stable set which facilitates control but also inhibits the spontaneity of that side as compared to the nondominant side. Jasper (1937, p. 161) specifically stated "... the tendency for the nondominant side to lead in attempted simultaneous movement may indicate a greater cortical control ('inhibition') of the movements of the so-called dominant side which is only a counterpart of the more highly perfected coordination of movement on this side." If so, the right side could become associated with control; whereas the left would betoken spontaneity. Thus, the relationship between poor heterosexual adjustment and focus on the right could be construed as indicating that those having difficulties in heterosexual expression are also those who "ignore" the spontaneous side of their bodies and concentrate on the "controlled" side. To attend to the right could represent a set to respond in a careful, self-controlled fashion, and such a set might be antithetical to the spontaneity required for adequate heterosexuality.

A second possible explanation for the relationship between focus on the right and heterosexual role relates to the fact that the right-handed individual is aware that his right hand is stronger than his left. He might, therefore, associate the right hand with strength and power which are attributes that typically define masculinity. But the left hand would be for him the "weaker one" and in that sense less masculine or more feminine. If one were doubtful about his masculine adequacy, he might express such concern in an anxious awareness of his right side which he equates with the strength needed to be masculine. He could be thought of as anxiously watching his right side because he anticipates it will not function to provide the power he feels he needs to be manly.

FRONT-BACK

Another major body-image dimension which was studied relates to the differentiation between the front and the back of the body. It has been widely considered psychoanalytically, but little experimentally. Freud (1924, 1938), Abraham (1927), Ferenczi (1955), Fenichel (1945), Tausk (1933), and Schilder (1935) theorized that the back of one's body is largely associated with anal functions. Of course, Freud originated the idea. He developed the concept of an "anal personality" who is unconsciously preoccupied with anal sensations (linked with the back of the body) as the result of conflicts experienced during the period of childhood when control of the anal sphincter is learned. He proposed that the conflicts faced by the child during the anal period center on issues of obedience and passivity versus opposition and self assertion. At a more elementary level, they presumably relate to control versus lack of control of a body function which is considered to be dirty and socially unacceptable. The "anal personality" is portrayed as having great anxiety about the potential loss of control of his anal sphincter and the associated implications of disobedience and soiling aggression. He is therefore said to be defensively strict with himself about being spontaneous or impulsive. Also, he is defensively clean, orderly, and obedient. But it is theorized that while he exercises such restraint over himself he has an underlying resentment about being controlled which permeates his behavior in the form of negativism and stubbornness.

There have been efforts to study the "anal character" concept empirically. Questionnaires and projective tests have been used to evaluate the meaningfulness of "anality" as a trait. Beloff (1957), Barnes (1952), Krout and Tabin (1954), and Couch and Keniston (1960) have shown that questionnaire items presumably sampling anal attitudes can be formulated which are coherent statistically and also in relation to the Freudian "anal stage" model. Blum (1949) and Miller and Stine (1951) have findings indicating that projective responses to pictures and story

completions can be reliably analysed for anal themes.

The present study conjectured that the greater an individual's awareness of the back of his body the more he is concerned with anal sensations and therefore typical of the "anal character." Thus, it was hypothesized:

1. The more attention a man devotes to his back as compared to the front of his body the greater is his anxiety about impulses "spilling out" and the less does he manifest spontaneous, impulsive behavior.

2. The greater his back awareness the stronger is his tendency to avoid direct aggressive expression and instead to make use of negativism.

3. His level of anxiety when confronted with stimuli that refer to anal functioning is positively correlated with his degree of back focus.

4. The more he recalls his parents as providing a model of behavior minimizing the direct expression of aggressive impulses, the more intense is his back awareness.

This last hypothesis follows from the assumption that a source of the "anal character's" anxiety about his aggressive potentialities is his perception of his parents as disapproving of aggression by restricting its appearance in their own behavior and that of other family members.

5. A fifth hypothesis was derived from work by Miller and Stine (1951) in which it was observed that children whose fantasies were typified by anal themes were, in terms of sociometric criteria, unusually popular with their peers. Miller and Stine speculated that the controlled traits of the "anal character" might impress others as a sign of being steadfast and orderly and elicit favorable evaluations. Relatedly, Couch and Keniston (1960) concluded from their data that for the "anal retentive": "The necessity of friction and aggressiveness in competitive situations is strongly denied, and replaced consciously by reactive trust and tolerance for others [p. 172]." Thus, from two different perspectives there seemed to be evidence that the anal character is proficient in pleasing rather than antagonizing others in group settings. It was therefore hypothesized that degree of back awareness would be positively correlated with the individual's interest in group participation and also his popularity in such group situations.

Study 2A

The hypothesis predicting an inverse relation between degree of focus on one's back and behavioral spontaneity was the first to be tested. It should be noted that Couch and Keniston (1960) have shown that among a cluster of traits characterizing the "retentive anal character" self-control and restraints upon impulsive expression are the most prominent.

Procedure. The intensity of a subject's attention to his back was measured with the Body Focus Questionnaire (BFQ). Embedded in the BFQ form were references to six[2] paired front-back body sites (e.g., front of head versus back of head, front of neck versus back of neck), and the subject indicated in each case whether he was more aware of the front or the back site. His score could range from 0 through 6. It has been shown in a group of 52 male subjects that there is a correlation of .44 ($p < .01$) between test-retest BFQ Back scores secured with an intervening period of 1 week.

The Impulsive scale, one of seven contained in the Thurstone Temperament Schedule, was chosen to ascertain how much spontaneity typified the subject. Thurstone (1953) states, "High scores in this category indicate a happy-go-lucky, daredevil, carefree, acting-on-the-spur-of-the-moment disposition [p. 1]." The Impulse score is based on the subject's reports concerning his own behavior, as indicated by responding Yes, No, or ? to a series of statements.

In dealing with the spontaneity variable, an unpublished Anal Orderliness scale developed by Henry Murray was also administered to one sample. This scale contains 10 items which inquire concerning compulsive and perfectionistic behavior (e.g., "I do things more slowly and carefully than others"; "I am generally methodical and systematic in the way I go about things"). The subject indicates his degree of agreement on a 5-point scale.

Subjects. Three different samples of subjects were studied with the Thurstone scale. They consisted, respectively, of 40, 51, and 60 male college students. The median age in each of the groups was 20. A fourth sample of 52 students (median age 20) was studied with the Murray Anal Orderliness Scale.

---

[2] The limited number of back items is due to the limited number of homologous front-back sites on the body which can be clearly defined verbally for subjects.

Results. The BFQ Back median in Sample 1 was 2 (range 0 through 6). The Impulse median was 9 (range 5 through 17). When the trichotomized Back scores were related to the dichotomized (at the median) Impulse scores by means of chi-square, it was found that they were negatively linked at a borderline level ($\chi^2 = 4.6$, $df = 2$, $p = .10$).

In Sample 2 the mean BFQ Back score was 3.1 ($\sigma = 1.8$). For the Impulse scores the mean was 10.6 ($\sigma = 2.7$). A product-moment correlation of $-.24$ was found between the Back and Impulsive scores. With a one-tailed test, which was used because this was a cross-validation attempt, the coefficient is significant at the .05 level.

In Sample 3 the mean Back score was 2.5 ($\sigma = 1.7$). The mean Impulse score was 10.6 ($\sigma = 3.5$). A significant negative correlation of $-.26$ ($p < .05$, one-tailed test) was found between the two sets of scores.

The results for Sample 4, in which the Murray Anal Orderliness scale was employed, indicated that orderliness was significantly and positively correlated with BFQ Back ($r = .36$, $p < .01$, two-tail test), as predicted. The mean Orderliness score was 16.2 ($\sigma = 5.6$).

Discussion. The data from the four samples go along with the expectation that the greater a man's awareness of his back the more likely he is to avoid responses which are not carefully controlled. None of the individual relationships are large but their consistent directionality over four samples is encouraging. Since anxiety about loss of impulse control is a prominent difficulty ascribed to the "anal personality," the above finding adds weight to the notion that attention to one's back and "anality" have overlapping significance.

Study 2B

Other studies were undertaken to determine whether the relationships of BFQ Back to several other anal trait variables were in the predicted direction.

Procedure. A second hypothesis stated that back awareness would be negatively correlated with open aggressiveness and positively so with stubbornness or negativism. Two subscales of the Buss-Durkee Inventory in Buss (1961) were employed to get at the anger variables: (*a*) A 10-item Aggression scale which is typified by assertions like "Whoever insults me or my family is asking for it," "If I have to resort to physical violence to defend my right, I will." The subject responds by answering Yes or No to each item. (*b*) A 5-item Negativism scale which is represented by a statement like "When someone is bossy, I do the opposite of what he asks."

Back awareness was measured in this case and in relation to the other hypotheses that follow by means of the BFQ.

The hypothesis that back awareness would be positively correlated with anxiety about stimuli with anal significance was explored via responses to the Blacky Pictures. One Blacky picture is labeled Anal Sadism and depicts the dog named Blacky in a position where his anus is visible and it is apparent that he has just defecated. Response to this picture was measured with the same procedure that was used in determining response to the pictures with sexual content in the studies concerned with BFQ Right.

Another hypothesis had predicted that BFQ Back would be negatively correlated with the subject's perception of how openly his parents expressed anger. He indicated on a 3-point scale the degree to which each of 10 statements concerned with behavior expressive of anger applied first to his mother and then to his father. Examples of the statements follow: Likes a fight; Expresses anger openly and directly; Good at telling people off. The responses "Not at all true"; "Slightly true"; "Very true" were weighted, respectively, 0, 1, 2. Total scores could range from 0 through 20.

Subjects. The subjects were 55 male college students recruited by payment of a fee. Their median age was 20.

Results. The BFQ Back score mean for the subjects who completed the Buss-Durkee scales was 2.4 ($\sigma = 1.6$). The mean for the Buss-Durkee Aggression scale was 5.2 ($\sigma = 2.4$); and for the Negativism scale it was 1.9 ($\sigma = 1.4$). One can see in Table 5 that the BFQ Back has chance relation to Aggression, but that it is significantly correlated with Negativism in the predicted direction ($r = .27$, $p < .05$). Apparently, back awareness is not correlated with self-reports of overt aggression, but it is positively so with such reports of negativistic behavior.

The Blacky Pictures data were supportive of the proposition that back awareness is positively correlated with the level of anxiety aroused by representations of anal function. Table 5 shows that BFQ Back is correlated .26 ($p < .10$) with the

TABLE 5

PRODUCT-MOMENT CORRELATIONS OF BFQ BACK WITH "ANAL CHARACTER" VARIABLES

BFQ Back versus	r	Significance level
Buss-Durkee aggression	$-.13$ ($N = 52$)[a, b]	n.s.[b]
Buss-Durkee negativism	.27 ($N = 52$)	$< .05$
Blacky anal sadism	.26 ($N = 51$)	$< .10$
Father anger	$-.38$ ($N = 50$)	$< .01$
Mother anger	.06 ($N = 52$)	n.s.
Total affiliation with organizations	.21 ($N = 55$)	$> .10$
Total elective offices held	.27 ($N = 55$)	$< .05$

[a] N varies because the subjects gave incomplete responses to some procedures or else misunderstood the instructions. In the case of parental ratings there were instances in which a father or mother had died when the subject was still a young child and therefore could not be recalled.

[b] Does not even attain .20 level.

rank-order placement of the Anal Sadism picture. The higher the BFQ Back score the less willing was the subject to compose a story about the Anal Sadism picture. This borderline relationship was examined by means of chi-square in which the dichotomized (at median) Back scores were related to the trichotomized (equal thirds as possible) Blacky scores. A significant chi-square of 6.0 ($p = .05$, $df = 2$) was found.

One notes that the mean anger score attributed to mother was 7.9 ($\sigma = 3.7$) and to father 5.6 ($\sigma = 2.6$). Table 5 indicates that BFQ Back was, as predicted, negatively correlated with Father Anger ($r = -.38$, $p < .01$). It had only a chance correlation with Mother Anger. The hypothesis was supported in terms of father's recalled traits but not in terms of mother's.

Procedure. The fifth hypothesis proposed that BFQ Back would be positively correlated with degree of participation in group activities and also one's popularity in such groups. To obtain an index of the subject's amount of participation in groups, he was asked to list the organizations to which he had belonged in high school and his first year in college. His popularity in these groups was estimated by asking him to list the elective offices he had held in each.

The BFQ score in this study was based on the enlarged number of 19 items.

Subjects. The subjects were 55 male college students whose median age was 20.

Results. The data dealing with the relation of BFQ Back to group participation are mildly favorable to the proposed hypothesis. There is a trend for the predicted positive correlation between BFQ Back and Total Affiliation with Organizations, although it is not significant ($r = .21$, $p > .10$). Further, the relationship between BFQ Back and Total Elective Offices Held is significantly positive, as predicted ($r = .27$, $p < .05$). One can say that those with relatively greater back awareness are those who most frequently report election to office in the organizations to which they belonged.

Discussion. How has the concept of a link between back awareness and "anal character" traits fared? The results are encouraging. The most pinpointed evidence of anal involvement in back awareness is offered by the fact that BFQ Back is positively correlated with the subject's reluctance to deal with the Blacky Anal Sadism picture. Especially encouraging, too, are the Thurstone Impulse scale data and the Murray Anal Orderliness scale findings which indicate that the individual who focuses attention on his back is also one who restricts impulse expression and behaves with compulsive care. The related assumption that such an individual would also avoid direct expression of aggression and rely instead on indirect forms of stubbornness was only partially confirmed. BFQ Back turned out not to be correlated with Buss-Durkee Aggression, but positively so with Negativism. There was partial confirmation of the hypothesis that back awareness is negatively correlated with the degree to which one's parents are recalled as openly showing anger. The confirmation was only partial because, while the data involving recall of the

father's behavior were congruent with expectation, those pertaining to mother were not. It is possible that the mother's style of anger expression is less important than father's in providing a model for a son.

The most tangential prediction made assumed that an individual's degree of back focus would be positively linked with how involved he was in group activities and also how popular he was in such groups. The results affirmed the expectation about back awareness and popularity in organized groups but indicated only a nonsignificant trend in the predicted direction for group participation. The significant finding should be cautiously interpreted because subjects' reports of their own group behavior were used rather than more objective observations. However, it is also true that the significant finding is supported by the work of Miller and Stine (1951) in which preoccupation with anal themes in children was found to be related to their group popularity. The anal orientation presumably basic to back awareness does seem to make for popularity with one's peers. This could be regarded as a function of modulated self-control or "reactive trust and tolerance for others [Couch & Keniston, 1960]" which might have a pleasing conciliatory effect. One would have to guess too that the negativism usually ascribed to an anal orientation is not prominent in peer interactions. Perhaps such negativism is more common in encounters with authority figures.

Paranoid defense. Freud (1950) and other analytic theorists (e.g., Ferenczi, 1916) underscored the importance of "anal fixation" in the formation of the paranoid delusion. It was considered that the paranoid delusion represents an attempt to disown and project outwardly passive-receptive (homosexual) incorporative aims derived from fixation on anal-erogenous zones. Tausk (1933), Starcke (1920), and Van Ophuijsen (1920) attempted to demonstrate that the persecutor in the delusion is assigned attributes associated with anal sensations and the buttocks. Aronson (1952) and Meketon, Griffith, Taylor, and Wiedeman (1962) have tested the concept of paranoia as a defense against passive-feminine homosexual impulses by comparing the frequency of "homosexual signs" in the Rorschach responses of paranoid as contrasted to nonparanoid schizophrenics. The results have largely supported the concept. Moore and Selzer (1963) have shown that in terms of clinical reports homosexual conflicts are more prominent in the paranoid than the nonparanoid schizophrenic. Both clinical and experimental observations tend to concur with the psychoanalytic formulation that the paranoid system is a defense against homosexual fantasies linked with passive anal-receptive attitudes.

Study 2C

If the paranoid delusion is correlated with anxiety about fantasies with anal reference, it should follow from the front-back awareness work which has been described that the paranoid would have relatively high awareness of his back. That is, disturbance about anal issues would be accompanied by intensified back concern. Operationally, this means that the paranoid schizophrenic should have greater back awareness than the nonparanoid schizophrenic.

Procedure. Back awareness was measured with a 19-item front-back subscale imbedded in a Body Focus Questionnaire containing 110 items. Subjects were seen individually. An observer rated on a 3-point scale their level of cooperation.

Subjects. Forty-four male schizophrenics were evaluated (paranoid, 27; nonparanoid, 17). They had not received shock therapy up to at least 6 months prior to the test session. The median ages in the paranoid and nonparanoid groups were, respectively, 33 (range 22–47) and 35 (range 18–43). This difference is not significant. In both groups the median years of education was 12. Ratings for cooperation were not significantly different for the two groups. Equal proportions (67%) of each group were receiving tranquilizing medication; and the dosage levels were not significantly different.

Results. The mean BFQ Back score for the paranoids was 8.7 ($\sigma = 3.2$) and for the nonparanoids 6.5 ($\sigma = 3.4$). A t test indicated that, as predicted, the difference between the group was significant at the $< .001$ level ($t = 4.0$).

Discussion. The paranoids proved, in agreement with the hypothesis, to be more focused upon their backs than the nonpara-

noids. This presumably means that the schizophrenics most invested in defending themselves against anxiety-provoking anal impulses are also the most back aware. In predicting the elevated back awareness of the paranoid schizophrenic, the meaningfulness of the front-back body awareness dimension is further extended. More importantly, the findings support Freud's model concerning the relationship of anal sensations and anxieties to paranoia.

Front-back skin resistance ratio. The meaningfulness of the front-back body awareness dimension provided an opportunity for further testing of a theory regarding the relation between body perception and physiological activation. Fisher and Cleveland (1958) proposed that the more salient one body sector is in the body scheme as compared to another the relatively greater will be the physiological activation of the former. Basic to this formulation is the idea that there is a mutually reinforcing interaction between degree of attention given to a body sector and its level of activation. It is assumed that certain needs or anxieties may cause an individual to focus his attention persistently upon a body area and that such attention may produce an increment in activation of the area in the same way that thinking of certain muscles may result in an increase in their action potential. Or in the way that thinking about putting food into one's stomach may produce changes in stomach activation. In turn, the increased activation of an area results in it becoming a source of augmented sensory experience which draws further attention to it. Thus, a circular feedback system involving attention and activation level could be established. Such a system might be found to apply to various organs or types of tissue (e.g., skin, vasculature, heart). Support for this view has already come from previous studies in which relative salience, as defined by the size attributed to body parts in projective settings, proved to be correlated with their relative activation as represented by skin resistance levels. Relationships between relative attributed size and skin resistance have been demonstrated for the following body sectors: front versus back, right versus left, head versus trunk, upper half versus lower half (Fisher, 1958; Fisher, 1960a; Fisher, 1961a; Fisher, 1961b).

Study 2D

It should follow from the above formulation that the greater an individual's awareness of the back as contrasted to the front of his body the relatively more activated should be the first in relation to the second. It was hypothesized that the higher the BFQ Back score the lower would be the resistance level of the back (i.e., more activated) in relation to that of the front.

Procedure. Back awareness was measured by means of nine BFQ items. Skin resistance measures were recorded with a Brush direct writing oscillograph. There was a constant current supply of 20 microamperes and a DC amplifier for measuring the voltage across the subject. The record was calibrated in ohms. Separate balanced systems were utilized for the front measure and the back measure. Area of recording from the sites was equalized by means of pieces of tape with two holes, each ¼ inch in diameter, punched into them. The period of recordings was based on the time required for both sites to stabilize to a point of no change for a 15-second period. A minimum of 30 seconds was taken in any case. The median length of recording was 165 seconds. The front recording area was taken from the first flat surface on the neck just below the Adams apple. An homologous area on the back of the neck was selected as the back site. Recordings were taken from the neck because it is the only sector not covered by clothes in which homologous front and back sites can be selected with some accuracy. A final reactivity value was tabulated equal to the ratio of the front resistance level to the back resistance level (front resistance/back resistance) at the point of stabilization.[3] The larger the ratio the greater is the reactivity of the back site relative to that of the front.

Subjects. The subjects were 52 male college students with a median age of 19.

Results. The median BFQ Back score was 2. The median front-back skin re-

---

[3] The representativeness of the front-back reactivity ratio derived from the neck was evaluated in a special sample of 22 men. Front-back resistance readings were taken simultaneously from neck, upper chest region and lower chest region sites. The neck front-back resistance ratio proved to be correlated .43 ($p < .05$) with the upper chest front-back ratio and .42 ($p < .05$) with the lower chest front-back ratio. Thus, the front-back values from the neck are significantly related to those derived from other body sites.

TABLE 6

CHI-SQUARE ANALYSIS OF RELATIONSHIP BETWEEN BFQ BACK SCORES AND FRONT-BACK SKIN RESISTANCE RATIOS

		BFQ Back		
	High[a]	Low	χ^2	Significance level
High	19	7		
Skin resistance ratio			3.9**	< .05
Low	12	14		

[a] High = Above median. Low = At median or below.

sistance reactivity ratio was 1.3. Because of the skewed character of the distributions, chi-square was used to examine the relationship between BFQ Back and the front-back skin resistance ratio. One can see in Table 6 that, as predicted, they are positively and significantly interrelated ($\chi^2 = 3.9, p < .05$).

Discussion. Apparently, the greater the individual's relative focus of attention upon the back the greater is the activation of the skin of his back in relation to that of the front. As earlier indicated, similar types of relationships have been observed between the skin resistance levels of body areas and the relative sizes attributed to them. The present findings are particularly significant because they are the first to show a link between a direct appraisal of how much attention an individual focuses on a body sector and the activation level of the skin in that sector. As data accumulate, it becomes evident that the body image and distribution of excitation in the body are intimately interwoven.

The question still remains as to the degree to which the correlation between BFQ Back and the front-back resistance ratio reflects the fact that those with higher activation of the back are receiving a greater amount of sensory stimulation from the back which is derivative of the physiological activation. However, if one considers that intensity of back awareness has turned out to be related in a meaningful way to the "anal character" typology it becomes difficult to interpret such awareness as a simple expression of level of back activation. That is, some proportion of the back awareness would seem to be related to attitudes one has learned to take toward one's own anal regions and functions. How such attitudes may excite physiological activation of the back or in turn be intensified by physiological variables remains to be seen.

EYES

The role of the eyes in obtaining information and also in the expressive function of the face has stimulated speculation about their psychological and symbolic significance. It has been variously suggested that they are unconsciously equated with oral "taking in" processes; hostile ("evil eye") intent; wishes to see forbidden sexual scenes; and even the genital organ itself (Fenichel, 1945). It is possible to conceptualize most of these speculations within the category of incorporative intent. Whether they suggest oral, sexual, hostile, or nonhostile aims they depict the eyes as accomplishing these aims by functioning as a channel for admission and "taking in." It was therefore conjectured that eye awareness would be linked with incorporative attitudes. But since the eyes are not truly incorporative in the way that a body opening like the mouth is, it was considered that the person who focuses on his eyes is probably one who is fearful of real incorporative wishes and therefore substitutes the sort of unreal ones exemplified by defining the eye as an oral channel. This view led to the hypothesis that degree of eye awareness is positively correlated with anxiety about incorporation and negatively so with indicators of free expression of incorporative wishes. The model for this formulation is provided by the Freudian concept that the use of a substitute zone for an erogenous purpose is due to anxiety which prevents use of the corresponding real erogenous zone. Specific hypotheses which were derived are listed below:

1. Degree of eye interest is inversely related to the enjoyment of eating. The greater the emphasis upon the eyes (presumably as a substitute incorporative channel) the less does the primary oral zone serve as a source of pleasure.

2. Also, one would expect that the greater an individual's focus upon his eyes the more anxiety he would evidence when responding to food related stimuli.

3. Quite analogously, degree of eye focus should be positively correlated with amount of anxiety evoked by symbolic references to incorporation.

4. Finally, it seemed logical to expect that eye interest would be negatively correlated with the degree to which one's parents were recalled as generous and giving. If eye interest depicts a sense of not being able to secure oral gratification, one might anticipate that such an attitude would reflect experiences with parents who appeared to be selfish and unwilling to give of their resources.

Study 3

Procedure. Eye awareness was measured with the Body Focus Questionnaire (BFQ). Included in the BFQ array of paired-comparisons were 11 items in which the eyes were compared to other facial areas (e.g., eyes versus ears, eyes versus mouth, eye versus chin). The BFQ Eye score could range from 0 through 11. It has been found that the test-retest coefficient for Eye scores, with a week intervening, is .54 in a group of 52 subjects ($p < .001$).

The Byrne Food Attitude Scale (Byrne, Golightly, & Capaldi, 1963) was used to test the hypothesis that BFQ Eyes would be negatively correlated with enjoyment of eating. This scale is composed of 221 items which inquire concerning liking for foods, pleasantness associated with past eating experiences, cooking skill of mother, and importance of food as a reward and comfort. Responses to each item are registered by the subject in terms of True or False. Only 47 of the 221 items have shown scale coherence for males and these are the items which are scored. The higher the score the more the subject is considered to have a positive attitude toward eating.

To investigate whether BFQ Eyes is positively related to anxiety when confronted with food stimuli, a selective memory procedure was used which has proven successful in other studies (Fisher, 1964a, 1964b). This procedure assumes that if a subject is asked to learn anxiety-arousing material, his recall for it will be relatively poorer than for equated material without anxiety connotations. Subjects (in small groups) were asked to view for 1 minute a list of 20 words projected on a screen; and they were subsequently given 5 minutes to write down as many of the words as they could recall. The list consisted of 10 words referring to food and 10 without food implications which were of the same average length and randomly distributed. The words in the list are enumerated below:

Plan	Beef
Mint	Road
Hall	Pair
Bun	Broth
View	Cream
Raisin	Fair
Check	Tea
Honey	Trace
Book	Plum
Hash	Stone

A subject's score equalled the number of food words minus the number of nonfood words recalled.

To ascertain whether BFQ Eyes is positively correlated with the level of anxiety aroused by symbolic references to incorporation, the Blacky Pictures technique was again employed. Only the Oral Eroticism and Oral Sadism pictures, which respectively depict Blacky sucking mother's breast and biting mother's collar, were considered to be pertinent to the hypothesis. It was expected that BFQ Eyes would be positively correlated with the subject's reluctance to tell a story about each of these pictures.

The question whether BFQ Eyes would prove to be related to the subject's recall of his parents' generosity required that ratings of the parents be obtained. Each subject indicated on a 3-point scale how applicable to his mother were each of nine statements concerned with generosity (e.g., "Feels we should help those weaker than ourselves"; "Considers it important to help charitable causes"). The same responses were obtained with regard to father. Weights of 0, 1, and 2 were applied respectively to the response alternatives of "Not at all true," "Slightly true," and "Very true." Total scores could range from 0 through 18.

SUBJECTS. The subjects consisted of 62 male college students recruited by payment of a fee. Their median age was 20.

Results. The mean BFQ Eyes score was 6.4 ($\sigma = 2.5$). The mean Byrne Food Attitude score was 34.5 ($\sigma = 5.5$). Table 7 indicates that BFQ Eyes is, as predicted, significantly and negatively correlated with the Byrne index ($r = -.29$, $p = .03$). The greater the subject's focus on his eyes the less he reports enjoyment of food-related experiences.

The mean selective food memory score was $+.2$ ($\sigma = 2.3$), indicating only a slight tendency for the group to remember more food than nonfood words. A correlation of $-.29$ ($p < .05$) was found between BFQ Eyes and the memory score. It would appear, as hypothesized, that with increasing awareness of one's eyes there is a parallel tendency to show selectively poor recall for references to food.

TABLE 7

PRODUCT-MOMENT CORRELATIONS OF BFQ EYES WITH INCORPORATIVE ATTITUDE VARIABLES

BFQ Eyes versus	r	Significance level
Byrne food attitude score	$-.29$ ($N = 61$)	.03
Selective food memory	$-.29$ ($N = 59$)[a]	$<.05$
Blacky oral eroticism	.17 ($N = 60$)	n.s.[b]
Blacky oral sadism	.28 ($N = 60$)	$<.05$
Father generosity	$-.30$ ($N = 57$)	.03
Mother generosity	$-.21$ ($N = 60$)	.10

[a] There are variations in N because some protocols had to be discarded either as a consequence of misunderstandings of instructions or the inappropriateness of certain questions (e.g., pertaining to a father or a mother who was long deceased).

[b] Does not even attain .20 level of significance.

The results for the Blacky Pictures were not as clearcut. The mean Oral Eroticism rank was 3.0 ($\sigma = 7.2$), and the mean Oral Sadism rank was 6.7 ($\sigma = 2.4$). Table 7 indicates that while BFQ Eyes was correlated in the predicted direction with Oral Eroticism ($r = .17$), it was not significantly so. However, BFQ Eyes was significantly correlated in the expected direction with Oral Sadism ($r = .28$, $p < .05$). These findings modestly support the view that the more aware an individual is of his eyes the greater his anxiety when perceiving oral themes.

The data involving the ratings of parental generosity indicated a mean of 5.1 ($\sigma = 2.0$) for Father and a mean of 11.7 ($\sigma = 3.4$) for Mother. Table 7 reveals that BFQ Eyes has a significant correlation of $-.30$ ($p = .03$) with Father Generosity, but a less impressive correlation of $-.21$ ($p = .10$) with Mother Generosity.

Discussion. Once again a psychoanalytic framework has bridged the gap between findings which concern body sensations and those depicting personality patterns. Eye awareness has proven to be related to anxiety about eating and food as indicated by the results involving the Byrne Food Attitude Scale and the memory for food words. In a more modest way, the Blacky Pictures findings suggest that eye awareness may also be associated with anxiety about incorporation defined in a general symbolic sense. Evidence was found too that eye awareness is linked with a male subject's recall of the generosity of his father, but not with his recall of mother's generosity. The predictions which were supported evolved from the complex assumption that if an individual concentrates his attention upon a body area capable of serving as a substitute or symbolic opening he does so because he is fearful of experiences with some primary body opening. Presumably, the focus upon the symbolic opening is an indirect attempt to experience what is forbidden elsewhere in the body. This assumption is a derivative of Breuer and Freud's generalized theory concerning the mechanisms underlying conversion phenomena.

BODY AWARENESS

The way a person distributes his attention to his body may be conceptualized not only in terms of the amount he focuses upon various body regions, but also with regard to the relative amount he gives to his own body as compared to other objects in his environs. Previous studies indicate there is great individual variation in the attention devoted to one's own body. Some are intensely concerned with their own body sensations; and at the opposite extreme are others who have little such awareness. Measurement of body awareness has proven to be feasible with a technique based on the frequency with which an individual refers to his own body when a sample is taken of what lies within his immediate awareness (Fisher, 1964b). With this technique it has been possible to demonstrate that in males there is a positive relationship between body awareness and the prominence of the nutritive-digestive areas in the body scheme. The more aware a man is of his body the more he focuses attention upon his stomach, gut, mouth, and related accessory sectors.

It has been shown that general body awareness is positively correlated with stomach awareness as defined by responses to paired body comparisons involving the stomach (e.g., stomach versus heart, stomach versus arms) in the BFQ. Further, body awareness has proven to be positively correlated with selective superior recall for words pertaining to oral-nutritive parts (e.g., mouth, stomach) as compared to words referring to nonnutritive areas (e.g., spine, skull) of the body (Fisher, 1964b).

The preoccupation with nutritive-digestive body sectors accompanying high body awareness in men intimated that sensations from the nutritive sectors must by their own salience and the sensations they arouse in other body systems play a large role in drawing the individual's attention to his body. If oral sensations contribute heavily to the male's body awareness, it is logical to expect that at still another level body awareness is related to incorporative attitudes. The presence of persistent sensations in oral-digestive regions could indicate anxiety about the incorporative functions of these regions. That is, it might be the person who has learned to be fearful about incorporation who becomes preoccupied with sensations from body sectors participating in incorporative activity and who, therefore, under the stimulus of tuning in on such sensations arrives at an unusual awareness of his body, as against other perceptual objects. The fact that body awareness is not correlated with awareness of other localized body regions besides the oral-nutritive ones indicates that such sensations have special potency in drawing the male's attention to his body.

Study 4

In view of the above findings, the following was hypothesized:

1. The greater an individual's awareness of his body the higher will be his underlying anxiety about incorporation and therefore the more limited his ability to enjoy the incorporative process exemplified in eating.

2. It was anticipated that general body awareness would be positively correlated with the amount of anxiety aroused by references to incorporative themes.

3. Further, if an individual's degree of body awareness derives from anxiety about incorporation one might expect that it would be negatively related to how altruistic he recalls his parents to have been. If there is anxiety about oral gratification, it could be a function of experiences with parents who were apparently unwilling to give.

Procedures. The prominence of the subject's body in his own perceptual field was measured in terms of what lay within his awareness at a given time. He was asked (in a group) to list on a sheet of paper "twenty things that you are aware of or conscious of right now." The 20 responses given were scored by summing the number of references he made to his own body. Such body references were defined so as to include explicit body designations (e.g., "My head hurts"), temperature or kinesthetic sensations, eating experiences (e.g., "I would like to eat a piece of pie") and descriptions of one's own clothing (e.g., "My shirt is blue"). Interscorer agreement for two judges for 59 protocols was 95%. The rationale for this measurement is that the greater an individual's perceptual focus upon his body the more should his body (or appropriate equivalents) find representation in his reports regarding the content of his awareness. A subject's score could range from 0 through 20.

The Byrne Food Attitude Scale, described above, was once again used to determine the subject's enjoyment and interest in eating. Also, a measure of his preferences for a list of 103 foods (20 of which are part of the Byrne scale) was obtained. He was asked to indicate for each food item whether he liked or disliked it. His score was the total number of foods liked. It could range from 0 through 103.

The Blacky Pictures ranking procedure was once again used to evaluate the subject's anxiety when responding to stimuli referring to incorporation. It was expected that the two Blacky pictures pertaining to orality (Oral Eroticism, Oral Sadism) would be given low preference by those with high body awareness.

The amount of altruism attributed to each of the parents was appraised with the same series of nine items used to measure parental altruism in the study described above of the BFQ eye variable.

Subjects. The subjects were 58 male college students (median age 20).

Results. The mean Body Prominence score was 3.7 ($\sigma = 2.5$). The mean Byrne Food Attitude Scale score was 34.8 ($\sigma = 5.4$).

Table 8 indicates that the Body Promi-

TABLE 8

PRODUCT-MOMENT CORRELATIONS OF BODY AWARENESS WITH INDEXES RELATING TO INCORPORATION

Body awareness versus	r	Significance level
Byrne food attitude scale	$-.30$ ($N = 58$)	$<.05$
Total number of foods liked	$-.25$ ($N = 58$)	.05
Blacky oral eroticism	.25 ($N = 57$)	$>.05$
Blacky oral sadism	$-.11$ ($N = 56$)	n.s.[a]
Father altruism	.03 ($N = 54$)	n.s.
Mother altruism	$-.11$ ($N = 57$)	n.s.

[a] Does not even attain .20 level.

nence score was, as predicted, negatively and significantly correlated with the Byrne Food Attitude Scale ($r = -.30$, $p < .05$). Included in the Byrne scale is a list of 103 foods, and the subject indicates which he likes and dislikes. The mean number of foods liked was 80.6 ($\sigma = 11.8$). Body Prominence proved to be negatively correlated with the number of foods liked ($r = -.25$, $p = .05$). These data indicate that the less pleasurable eating appears to an individual the greater his body awareness.

For the Blacky Pictures scores the mean rank for Oral Eroticism was 7.2 ($\sigma = 3.0$) and for Oral Sadism 6.8 ($\sigma = 2.5$). Table 8 indicates that Body Prominence was, as predicted, positively correlated with the degree to which the Oral Eroticism picture was put low in the rank sequence ($r = .25$). The p value for the correlation is just short of the .05 level. This relationship was examined further by means of a chi-square comparison in which the trichotomized (into as equal thirds as possible) Prominence scores were related to the dichotomized (at median) Blacky scores. The χ^2 of 6.8 ($df = 2$) was significant at the $<.05$ level. However, the prediction regarding the relationship of Body Prominence to Oral Sadism was not borne out.

The Father Altruism mean was 10.1 ($\sigma = 3.3$) and the Mother Altruism mean 11.8 ($\sigma = 3.5$). Table 8 indicates that Body Prominence was, contrary to prediction, not significantly related to these variables.

Discussion. There is only modest congruence between the data and the hypotheses. The best results were obtained for the prediction that Body Prominence would be inverse to satisfaction derived from eating. The Byrne scale and the index of number of foods liked were both related to Body Prominence in the fashion anticipated. With increased Body Prominence there is a corresponding negative attitude toward food intake which can be interpreted as relating to anxiety about incorporation.

It is encouraging too that with increasing Body Prominence one finds augmented anxiety about the Blacky Oral Eroticism picture as defined by unwillingness to relate a story about it. But Body Prominence did not turn out to be related to Oral Sadism. This may reflect the weakness of the hypothesis. It is also possible that the oral anxiety linked with body awareness pertains specifically to incorporation (as represented by the Oral Eroticism picture) and not to the sadistic, biting intent portrayed in the Oral Sadism picture.

The formulation relating Body Prominence, with its presumed concern about incorporation, to the subject's recall of the degree of selfishness of each of his parents was not affirmed by the findings. One cannot trace the attitude toward incorporation which is linked with body awareness to the simple matter of the parents' recalled generosity.

The results show promising continuity. If one considers that body awareness is associated with focusing upon the oral regions of one's body, inhibited eating behavior, and also anxiety in the perception of a pictured oral incorporative theme, it is clear that there is some substance to the idea that the amount of attention a man directs to his body is related to how much anxiety he has about taking in and con-

suming. Why such a relationship should exist is puzzling. However, it has been noted (Fisher, 1964b) that while body awareness is encouraged in woman by the culture it is discouraged for men. Van Lennep (1957) reported that the male manifests decreasing interest in his own body as he matures beyond adolescence. For the female the opposite is true. Perhaps the association in the male between body awareness and incorporative anxiety represents the fact that it is the male with oral problems who, in terms of the literature dealing with orality (Blum, 1949; Fenichel, 1945) would be expected to have difficulty in being an independent manly person, is also the one to be concerned in an unmanly way with his body sensations.

HEART

Unrealistic concern with one's heart has been reported often as a neurotic symptom (Fenichel, 1945; Schneider, 1954). It has been conjectured that such concern reflects factors like repressed sexual excitement, unexpressed rage, and fear of death. In scanning the statements in the literature about what characterizes the person who focuses upon his heart, one finds their diversity difficult to integrate. Little agreement exists as to which affects or impulses might preoccupy the heart-conscious individual. However, there are intriguing references to the idea that the heart, because of its special importance and its unique prominence as a source of body sensations and rhythms, may easily become involved with the individual's fantasies and conflicts. Perhaps it offers a convenient focus for feelings and anxieties about oneself.

The heart variable seemed to be worth study, but there was little material available from which to derive hypotheses about its possible personality relationships. Therefore, the decision was made to undertake, first of all, some general explorations by means of the Body Focus Questionnaire (BFQ), which has already been described. It contains a subscale of 16 items sampling how aware the individual is of his heart. The Heart subscale has shown a test-retest reliability of .62 ($N = 50$) over a period of 1 week. A number of widely scanning studies have been pursued with BFQ Heart to ascertain its relationships with personality measures (e.g., Edwards Preference Schedule) and social variables (e.g., social class, religion), but the results have been largely of a chance order. It would serve little purpose to describe them. One of the few promising trends that did emerge was the observation that an individual's heart awareness is positively related to his ratings of religiosity of his parents and also himself. This finding seemed noteworthy in light of previous reports that persons with anxiety about their hearts are unusually conscientious (Ross, 1945). In fact, a study by Wittkower, Rodger, and Wilson (1941) portrays such persons as puritanical, with a strong sense of duty and morality. Thus, one could discern an initial basis for regarding heart awareness as related to issues of morality, religiosity, and virtuous conformity. The possibility presented itself that heart awareness would be positively correlated with an approach to life emphasizing religious commitment, with its accompanying concern about issues of right and wrong.

Study 5

Using this framework, the following hypotheses were ventured:

1. The greater an individual's awareness of his heart the more religious should be his orientation.

2. A derived assumption is that his degree of heart focus would be positively related to the amount of religiosity he ascribes to his parents.

3. With increasing heart awareness there should be enhanced concern and guilt about wrongdoing.

Degree of heart focus should be positively linked with anxiety about sexual expression, since sexual behavior is among the most stringently regulated by religious standards.

Relatedly, it was anticipated that intensity of heart focus would be inverse to the amount of sexual expressiveness the individual recalls as typifying his parents.

A more tangential prediction was also made about the relationship between heart awareness and openness to aesthetic ex-

periences. Among the few significant relationships observed in earlier exploratory studies was a negative correlation between BFQ Heart and the Aesthetic subscale of the Allport-Vernon-Lindzey Study of Values. This suggested that the more aware an individual is of his heart the less is his interest in, and sensitivity to, artistic and imaginative representations. The pertinence of this hypothesis to the tentatively formulated religious-oriented picture of the heart-oriented person is pointed up by the fact that previous studies (Allen, 1955) reported a trend for religiosity and aesthetic interest to be negatively correlated.

Procedure. Heart awareness was measured with the 16-item Heart subscale of the BFQ.

Several procedures were employed to evaluate religiosity.

1. The subject estimated the average number of times per month he currently attended church.

2. He rated his own level of religiosity on a 5-point scale.

3. His score on the Religious subscale of the Allport-Vernon-Lindzey Study of Values was determined.

The religiosity ascribed by the subject to his parents was evaluated by obtaining his estimates of how often each, on the average, attends church per month. Also, he indicated on a 5-point scale his response to the following question: How important a part did religion play in your family when you were growing up?

Measurement of guilt and anxiety about wrongdoing was approached in two ways.

1. One involved a selective memory task. It was anticipated that the higher an individual's sense of guilt the more he would selectively forget words he had learned which referred to guilt themes. The following list of words was exposed on a screen in a group setting.

Round	Honest*
Fault*	Sight
Judge*	Happy
Across	Bible*
Book	Forge*
Steal*	Worker
Ready	Law*
Rule*	Paint
Bark	Wrong*
Verdict*	Clerk

[Guilt words are starred]

The subject was told to study the list. After he had done so for 1 minute, he was given 5 minutes to write on a sheet as many of the words as he could recall. Ten of the words in the list refer directly or indirectly to guilt linked ideas; and 10 are neutral. The mean length of the two sets of words is equivalent. A selective memory score was derived equal to the number of guilt words minus the number of nonguilt words recalled.

2. A second measure of guilt concern was derived from a previously described ranking procedure involving the Blacky Pictures. One of the Blacky Pictures, titled "Guilt Feelings," shows the dog Blacky being reproved by a figure symbolic of his conscience. It was predicted that the higher the BFQ Heart score the lower would be the rank assigned to the Guilt Feelings theme.

Four different procedures were utilized to examine the subject's orientation toward a sexual role and sexual expression.

1. Amount of heterosexual activity was taken as one criterion of freedom to be sexually expressive. It was appraised by means of the same, earlier described, questionnaire which inquires concerning frequency of dating in high school and college.

2. Selective memory for sexual words served as another index of anxiety about sexual issues. Subjects viewed the following lists of words for 1 minute and were then given 5 minutes to recall them.

Plan	Caress*
Touch*	Train
Debate	Listen
Run	Perfume*
Dance*	Write
Feel*	Kiss*
Build	Twist*
Flirt*	Skate
Dust	Hug*
Date*	Color

[Sexual words are starred]

Ten of the words have sexual connotations. The other 10 are neutral and of the same average length as the sexual ones. A memory score was computed equal to the number of sexual minus the number of nonsexual words recalled. It was considered that the greater the subject's anxiety about sexual matters the more he would show selectively poor recall for the sexual words.

3. A third way of sampling sexual anxiety involved the Blacky Pictures ranking procedure. The lower the ranks assigned by the subject to the three pictures with sex related themes (Love Object, Oedipal Intensity, and Castration Anxiety) the more elevated was his sexual anxiety taken to be.

4. Still another approach to the matter of sexual orientation was attempted by means of the Vague Sex Pictures earlier described. These are the pictures which, by virtue of their vague definition of the sex of the human figures shown, are intended to arouse anxiety in those who have poorly defined concepts of their own sex roles. Degree of anxiety aroused by these pictures is evaluated in terms of the amount of negative affect they evoke, as evidenced in ratings of the attractiveness and intelligence of the figures.

The subject's perception of how freely his parents expressed themselves with regard to sexual matters was tapped with an eight-item questionnaire. It was first requested that he indicate on a

3-point scale the degree to which such items as the following applied to his father and then to his mother:

1. Likes to be considered physically attractive by members of the opposite sex.
2. Tells jokes with sexual references.
3. Provides advice and counsel on sexual matters.

Weights of 0, 1, and 2 were applied respectively to the response alternatives of "Not at all true," "Slightly true," and "Very true." Total scores could range from 0 through 16.

Motivation for seeking out and opening oneself to aesthetic experiences was measured with the Aesthetic score of the Allport-Vernon-Lindzey Study of Values. This is the index which has already been mentioned as having been related to BFQ Heart in an earlier exploratory study.

Subjects. Sixty-one male students participated as subjects (median age 21).

Results. The mean BFQ Heart score was 5.3 ($\sigma = 4.3$). Table 9 indicates some support for the hypothesis that BFQ Heart is positively correlated with degree of religiosity. One finds BFQ Heart positively related at a borderline level with estimate of frequency of church attendance and also self-rating of religiosity. When church attendance frequency and self-rating of religiosity were simply summed for each subject, this combined index was significantly related to BFQ Heart ($\chi^2 = 6.9$ [$df = 1$], $p < .01$). A relationship of .40 ($p < .005$) was found between BFQ Heart and the Study of Values Religious score.

Table 9 demonstrates too that there are trends for the subject's BFQ Heart score to be positively linked with the level of religiosity he attributes to his family. It is positively correlated ($r = .38$, $p < .01$) with estimates of frequency of mother's church attendance and ratings of importance of religion in the family ($r = .24$, $p < .10$). While it is positively correlated with estimates of the frequency of father's church attendance, the coefficient is not significant.

The idea that BFQ Heart is tied in with a sense of guilt about wrongdoing was slightly reinforced by its borderline negative correlation ($r = -.22$, $p < .10$) with selective memory for words referring to guilt themes. This relationship was examined further by means of chi-square. The trichotomized (as equal thirds as possible) Heart scores were compared with the dichotomized (at median) memory score. A χ^2 of 9.1 ($df = 2$) was found which is significant at the <.02 level. However, skepticism is encouraged by the fact that BFQ Heart was not significantly related to the rank assigned to the Blacky Guilt Feeling picture.

The findings for the sexual behavior variables were equivocal. Table 11 shows that BFQ Heart has only a chance relationship to average number of dates per week in high school and college and also to the index of serious dating.

TABLE 9

PRODUCT-MOMENT CORRELATIONS OF BFQ HEART WITH RELIGIOUS VARIABLES

BFQ Heart versus	r	Significance level
Estimate of frequency of church attendance per month	.23 ($N = 61$)	<.10
Self-rating of religiosity	.24 ($N = 61$)	<.10
Study of Values religious score	.40 ($N = 58$)	<.01
Estimate of frequency of father's church attendance per month	.15 ($N = 56$)	n.s.
Estimate of frequency of mother's church attendance per month	.38 ($N = 60$)	<.01
Estimate of importance of religion in family	.24 ($N = 61$)	<.10

TABLE 10
PRODUCT-MOMENT CORRELATIONS OF BFQ HEART WITH GUILT INDEXES

BFQ Heart versus	r	Significance level
Guilt memory	$-.22$ ($N = 60$)	$<.10$
Blacky guilt feelings	$.12$ ($N = 60$)	n.s.

A borderline negative correlation in the predicted direction was observed between BFQ Heart and selective recall for sexual words ($r = -.23$, $p < .10$). Similarly, BFQ Heart related in the predicted direction at a borderline level to Blacky Castration Anxiety ($r = .24$, $p < .10$). This relationship was appraised also by means of chi-square involving the comparison of dichotomized (at median) Heart scores with trichotomized (equal thirds as possible) Blacky scores. The χ^2 was 9.7 ($df = 2$) which is significant at the $<.01$ level. BFQ Heart had a chance relationship to Blacky Oedipal Intensity and one with Blacky Love Object that was significant in a direction opposite to that predicted.

BFQ Heart was correlated .28 ($p < .05$) with the degree to which the Vague Sex Pictures were perceived as ugly. The correlation of BFQ Heart with how unfriendly the Vague Sex Picture figures were judged to be was not significant.

TABLE 11
PRODUCT-MOMENT CORRELATIONS OF BFQ HEART WITH SEXUAL INDEXES

	r	Significance level
Heterosexual behavior		
Average number of dates per month in high school	$.08$ ($N = 61$)	n.s.
Average number of dates per month in college	$-.02$ ($N = 58$)	n.s.
Index of serious dating	$-.03$ ($N = 60$)	n.s.
Sexual memory	$-.23$ ($N = 61$)	$<.10$
Blacky Pictures		
Blacky castration anxiety	$.24$ ($N = 60$)	$<.10$
Blacky oedipal intensity	$.18$ ($N = 60$)	n.s.
Blacky love object	$-.37$ ($N = 60$)	$<.01$
Parental sexual behavior		
Recall of father's sexual expressiveness	$-.04$ ($N = 57$)	n.s.
Recall of mother's sexual expressiveness	$-.42$ ($N = 59$)	$<.001$
Vague sex pictures		
Ugly ratings	$.28$ ($N = 56$)	$<.05$
Unfriendly ratings	$-.11$ ($N = 56$)	n.s.

As anticipated, BFQ Heart had a substantial negative correlation ($r = -.42$, $p < .001$) with recall of how sexually expressive mother was but a chance correlation with the same index as it applies to father.

A final result to be mentioned is the fact that there was an encouraging correlation of $-.31$ ($p < .03$) in the predicted direction between BFQ Heart and the Aesthetic scale of the Study of Values.

Discussion. The results embracing religiosity of self and recalled religiosity of one's parents concur with the hypothesis that the more aware an individual is of his heart the greater his current and past commitment to religious values. Although some of the correlations between BFQ Heart and religious parameters (e.g., frequency of church attendance and self-rating of religiosity) are not very substantial, they still carry weight because they represent cross-validation of the same relationships which were observed in an earlier study.

The hypothesis concerning the association of guilt about wrongdoing with heart awareness was supported by the findings involving selective memory for guilt words but not so by the Blacky Guilt Feelings data. The results for the selective memory variable do seem to be worth further study. They not only attained statistical significance, but have an attractive pertinence to the concept of heart awareness as being a function of a religious, moralistic orientation.

There is no sign of a relationship between BFQ Heart and reported frequency of heterosexual behavior. Heart awareness showed a borderline inverse relationship to the ability to recall sexual words that had been learned. If the repression of the sexual words is ascribed to their arousal of anxiety, one can consider the possibility that at least at the level of thought and verbal concept increased heart awareness might be accompanied by increased anxiety about sexual themes. The possibility that heart awareness is linked with sexual anxiety was also reinforced by the significant positive correlation found between BFQ Heart and ratings of ugliness of the Vague Sex Pictures. But a skeptical attitude is encouraged by the fact that the correlations of BFQ Heart and the Blacky Pictures portraying sexual themes were largely not as predicted. To further complicate matters, one notes that BFQ Heart had, as predicted, a substantial negative correlation with recall of mother's sexual expressiveness, although not with father's. The results for the sexual variables are complex and inconclusive. A few leads are promising which indicate some relationship between heart awareness and sexual anxiety. The results do not suggest that heart awareness is related to sexual behavior in any generalized sense.

Cross-validation was obtained of the original finding that heart awareness is inverse to interest in aesthetic experience. If one conceptualizes aesthetic interest as indicating openness to novel representations and fantasy productions, it would follow that the heart-focused individual tends to seal himself off from such stimuli. This finding can be used as a keynote to integrate much of the data. Focusing upon one's heart can be regarded as part of a way of life which revolves about a closed-off world defined by religious precept and perhaps also guilt. It may be an important part of this way of life to feel guilt and anxiety about certain forms of fantasy, particularly those expressing sexual wishes or the urge to do what is wrong.

It becomes an exciting matter to determine why an individual's intensity of attention to his heart should be linked with such an orientation. One could pursue Fenichel's suggestion that the heart because of its rhythmic pulsation and its growing larger–growing smaller qualities is easily associated with sexual experience. As such, devotion of attention to one's heart could represent an anxious concern with an organ symbolizing illicit excitement incompatible with a religious orientation. Of course, an argument against this formulation would be the fact that BFQ Heart had rather inconclusive relationships with the sexual variables which were studied.

Another speculation could go to the opposite extreme and suggest that the heart is one of the morally "safest" body or-

gans to which one can direct one's attention. There are no taboos about referring or attending to one's heart. This contrasts with the fact that overtones of sex, dirt, and other embarrassing topics apply to many other major body sectors (e.g., gut, genitals). Perhaps the individual raised in a moralistic atmosphere which contains taboos about looking at, or touching, "bad" body regions would find his heart one of the few safe allowable body experiences. In focusing upon his heart he could experience awareness of his body, but without showing interest in the "bad" side of himself.

PERCEPTUAL SELECTIVITY IN TERMS OF THE AMES THERENESS-THATNESS APPARATUS

Study 6

Having completed the above studies, it was decided to apply the results to making pinpointed predictions about the relationship between BFQ variables and perceptual selectivity. The question was whether body-attention parameters could be used to anticipate how subjects would react to pictures with varying themes presented in the Ames Thereness and Thatness Table (T-T) (Kilpatrick, 1952). The Ames T-T can be used to create an ambiguous perceptual situation in which the value or emotional significance of a stimulus (e.g., picture) can be determined in terms of size or distance characteristics ascribed to it (Hastings, 1952; Hastorf, 1950; Kilpatrick, 1952). It was anticipated that if an individual were asked to make a judgment in the T-T setting about a stimulus touching on the conflict associated with one of his body attention patterns he would display sensitivity to that stimulus. For example, if he had a high BFQ Right score he would be expected to register an exaggerated response to a heterosexual theme. Or if he had a high Body Prominence score he might demonstrate an accentuated reaction to an oral stimulus. The following were the predictions made:

1. BFQ Right will be positively correlated with accentuated response to pictures with heterosexual content (viz., a nude female).

2. BFQ Back should be positively correlated with exaggerated reaction to a picture with homosexual connotations (e.g., rear view of a nude male). An explanation is in order concerning the derivation of this hypothesis. The data which have been collected relating to the Back dimension indicate that back awareness is positively correlated with the occurrence of "anal character" traits. Such traits are, within the psychoanalytic model, basic to an orientation typified by ambivalent attitudes toward men such as are associated with homosexual conflicts. It was with this concept in mind that it was earlier predicted and verified that paranoid schizophrenics, whose principal conflicts are presumably homosexual in nature, would be typified by high back awareness.

3. It may be anticipated that BFQ Eyes will be positively related to indicators of augmented response to an oral theme (viz., picture of ice cream).

4. The prediction about BFQ Eyes would be expected to apply analogously to the Body Prominence dimension.

5. The prediction chosen for BFQ Heart was to the effect that it would be positively correlated with accentuated response to the theme of flagrant sexuality (viz., nude female). This hypothesis derives, of course, from the idea that an open display of sexuality is particularly disapproved in religious and puritanical systems.

Procedure. Measures relating to Back, Right, Eyes, and Heart awareness were obtained with the 110-item BFQ form. Body Prominence was appraised with the same procedure as already described above.

The Thereness-Thatness technique was employed for measuring perceptual response. It consists of two viewing tunnels which are side by side. The tunnel on the right, which is completely dark, contains no cues for distance and therefore none for size. The stimulus to which the subject responds is projected on a screen set up in this tunnel at a distance of 2 meters from him, and it is viewed monocularly. In the left tunnel, viewed binocularly, there are five lucite rods (each lighted by a 15-watt incandescent lamp) at 65-centimeter intervals. A Clason projector, on the right side of the apparatus and shielded from the subject's view, was used to

project the image of a picture on the screen in the tunnel on the right. This projector can alter the size of the projected image over a wide range without significantly changing its clarity or brightness. As the image size is increased the picture seems to move toward the subject, and as it is decreased it appears to move away. It is therefore possible to present the subject with a judgmental task which seems to involve the spatial placement of a picture but which actually revolves about altering its size on the screen. The experimental task was one in which the subject was asked to view (with his head in a headrest) a projected picture in the right-side tunnel and told that he could, by means of a knob, move it forward or backwards on a track in order to line it up with rods in the left-side tunnel. The instructions were as follows:

> You will be looking at various pictures of objects which you will see in front of you. On your left you will see some lighted rods. Your job will be to turn the knob with your right hand and make the object line up with the rod I name. I want you to move the picture back and forth until it is even with the rod I name.

The size setting made with the knob could be read from a pointer attached to the lens holder that moved as the subject turned the knob. A scale from 1 to 13 was used, with larger values indicating a larger image and by implication closer optical placement. The voltage on the bulb in the Clason projector was kept at a maximum reading of 120 volts by means of an auto transformer, thus controlling its 4,250 lumen output.

Six pictures were presented (front view of clothed male, front view of female nude, rhombus-shaped geometric figure, ice cream parfait, front view of clothed female, and rear view of male nude). They were all line drawings of the same height and width; and presented in the sequence just enumerated. Judgments of the pictures were obtained under six different conditions. Each picture was first presented at the apparent furthest position from the subject, and he was asked to line it up with the rod second closest to him. A second series of trials involved telling the subject to move the picture from the closest possible position to the position of the fourth rod. Thirdly, the picture was to be shifted from midway (halfway point on size scale) to the fifth rod. Fourth in sequence was the task of moving the picture from the apparent closest position to the fifth rod. Next, the picture was to be moved from midway to the second rod position. Finally, the subject manipulated the picture from the farthest position to the apparent position of the third rod.

Prior to the experiment the subjects were tested for visual acuity and astigmatism, respectively, by means of a Snellen chart and an astigma sunburst chart. Only those with 20–20 vision and no astigmatic defects went on to participate. Five minutes of dark adaptation were allowed before the T-T task. Following the T-T trials the subject was asked to recall the pictures he had seen. He then undertook in sequence the Body Prominence and BFQ tasks. At the end of the session he was again asked to recall the T-T pictures.

The mean size setting of each picture for the six trials was computed. Also, the mean rank (Rank 1 = largest or "closest" setting) of each picture in relation to the other five pictures in the series was determined. The number of pictures the subject forgot or described with error in the two recall tasks was tabulated. Since six errors were possible for each recall, scores could range from 0–12. The purpose of this index was to ascertain the degree to which the subject seemed to be dealing repressively with the themes in the T-T pictures.

The analysis of the data was complicated by issues of "defensive style" which have already been described in previous T-T studies. Shellow (1956) found that subjects manifesting anxious involvement in the T-T task made pictures relatively large (i.e., apparently closer to themselves). Those without such involvement made pictures relatively small (i.e., apparently farther away). Analogous results were obtained by Hastings (1952) who noted that insecurity was positively correlated with setting pictures relatively "close." Also, Kaufer (in Ittelson & Kutash, 1961) reported that persons characterized by an anxious "moving away" from people put emotional pictures "closer" to themselves. This contrasted with subjects typified by a "moving toward" others orientation who put such pictures "farther away." In terms of previous experiments by Ittelson (in Kilpatrick, 1952) and Hastorf (1950) it is known that a picture presented in the T-T apparatus which is more vivid than another requires a smaller or "farther away" setting in order to be lined up with a spatial reference point. It is therefore likely that the anxiously involved subjects who put specific T-T pictures relatively "close" to themselves do so because they defensively minimize their intensity. An evasive orientation under the T-T viewing conditions results in the pictures being set larger or "closer" because they appear subjectively less intense. The pictures require extra "magnification" to match the standard of how large one would expect them to be at a given distance.

The two memory tasks included in the present T-T procedure provided a means for determining whether the subject took a repressing attitude toward the T-T pictures. They made it possible to evaluate whether his anxiety was sufficiently intense to intrude repressively upon his cognitive functioning. The analysis of the T-T data was based on a separation of subjects into those manifesting repression in their recall and those dealing nonrepressively with the pictures. This approach was encouraged by exploratory observations indicating that the relationships between BFQ scores and T-T settings were frequently reversed in the two groups. Thus, as anticipated, specific BFQ scores in the repression group tended to be positively correlated with setting given pictures larger

(apparently closer); while in the nonrepression group the relations between such BFQ scores and T-T settings were in the opposite direction.

Subjects. The subjects consisted of 54 male college students. Their median age was 20.

Results. A division was made between subjects who evidenced no errors in their recall of the T-T pictures and those with two or more errors. Four subjects who made only one error were not included in the analysis in order to have a clear cutting point between the error and no error groups. This categorization derived from the formulation that a repressing (forgetting) response to the pictures should be expressed in a different mode of perceptual defense than a nonrepressing response. Twenty-five of the subjects proved to be Repressors and 29 Nonrepressors.

The results pertaining to each BFQ category will be considered in turn.

1. Right-Left. The mean BFQ Right score in the Repression group was 8.7 ($\sigma = 3.4$); and in the Nonrepression group it was 8.1 ($\sigma = 3.2$). The T-T index to which the Right scores were related involved the difference in rank between the female nude setting and the male nude setting. This index was chosen because it evaluates the degree to which the subject responds selectively to a heterosexual, as compared to a nonheterosexual, nudity theme. The coding of the rank difference scores was such that the more negative they were the larger (closer) was the female as compared to the male setting. In the Repression group the mean T-T difference score was $-.1$ ($\sigma = 3.0$); and in the Nonrepression group it was $-.8$ ($\sigma = 2.8$).

A significant chi-square ($\chi^2 = 6.8$, $df = 2$, $p < .05$) was found in the Repression group between BFQ Right and the T-T female nude rank–male nude rank difference. That is, the higher the subject's Right score the greater was his tendency to make the female nude picture larger (closer) than the male nude picture. The chi-square between BFQ Right and the female nude rank–male nude rank difference was not significant for the Nonrepression subjects.

2. Front-Back. Mean BFQ Back scores were respectively 7.2 ($\sigma = 3.5$) and 7.8 ($\sigma = 4.5$) in the Repression and Nonrepression groups.

The same female nude rank minus male nude rank T-T index was used as just described above. In the present instance it was intended to tap differential response

TABLE 12

CHI-SQUARE ANALYSIS OF SIGNIFICANT OR NEAR SIGNIFICANT RELATIONSHIPS OF BODY ATTENTION VARIABLES TO THERENESS-THATNESS INDEXES IN ANXIOUS GROUP

Variable	Thereness-Thatness				
		Female nude rank–Male nude rank			
		H	M	L^b	χ^2
BFQ Right	H^a	2	10	7	6.8**
	L	8	4	4	($df = 2$)
		Female nude rank–Male nude rank			
		H	M	L^b	
BFQ Back	H^a	7	3	2	7.4**
	L	1	6	6	($df = 2$)
		Sum of nude ranks–Parfait rank			
		H^a	L		
BFQ Eyes	H	7	3		
	M	6	3		5.3*
	L^b	1	5		($df = 2$)
		Female nude rank–Female clothed rank			
		H	L^b		
BFQ Heart	H^a	3	9		4.9**
	L	9	4		($df = 1$)

[a] Split at median.
[b] Split into as equal thirds as possible.
* $p < .10$.
** $p < .05$.

TABLE 13

CHI-SQUARE ANALYSIS OF SIGNIFICANT OR NEAR SIGNIFICANT RELATIONSHIPS OF BODY ATTENTION VARIABLES TO THERENESS-THATNESS INDEXES IN NON-ANXIOUS GROUP

		Sum of nude ranks–Parfait rank		
		H	L^a	χ^2
Body prominence	H	4	4	
	M	3	9	5.8*
	L^b	7	2	($df = 2$)
		H	L^a	
BFQ Heart	H	6	3	6.0**
	M	3	4	($df = 2$)
	L^b	2	11	

[a] Split at median.
[b] Split into as equal thirds as possible.
* $p < .10$.
** $p = .05$.

to the male (homosexual) as compared to the female theme. The means and standard deviations were the same as those cited for the right-left data.

In the Repression sample a significant chi-square was found between BFQ Back and the tendency to make the male nude larger (closer) than the female nude ($\chi^2 = 7.4$, $df = 2$, $p = .025$). The equivalent relationship in the Nonrepression sample was not significant.

3. Eyes. The mean BFQ Eye score was 7.2 ($\sigma = 2.9$) for the Repression subjects and 7.1 ($\sigma = 2.8$) for those in the Nonrepression category. The T-T index employed was the difference between the average of the ranks of the settings for the two nude figures minus the rank of the ice cream parfait setting. It was intended in this way to determine if the response to the oral stimulus was different from the response to the other two most vivid or ego-involving picture stimuli in the series. The mean difference score was $-.4$ ($\sigma = 2.6$) for the Repressors and -1.2 ($\sigma = 2.6$) for the Nonrepressors.

Table 12 indicates that there was in the Repression sample a borderline relationship at the $<.10$ level between BFQ Eyes and the difference between the average of the nude ranks and the parfait rank ($\chi^2 = 5.3$, $df = 2$, $p < .10$). The higher the subject's Eye score the greater the tendency to set the parfait picture relatively larger (closer) than the nude pictures.

The equivalent relationship in the Nonrepression group was of a completely chance order.

4. Body Prominence. Mean Prominence scores in the Repression and Nonrepression categories were respectively 2.5 ($\sigma = 2.0$) and 2.8 ($\sigma = 2.0$).

The same index for evaluating the response to the ice cream parfait was used as described above in the analysis of the BFQ Eye data. Means and standard deviations of the distributions were also the same. A chance relationship was observed in the Repression group between Prominence and the relative setting of the parfait picture. But the results in the Nonrepression sample (Table 13) indicated that the chi-square depicting the relation between Prominence and the parfait index was 5.8 (in the predicted direction), which is just short of the 6.0 needed for significance at the .05 level. The greater the subject's body awareness the smaller (further away) did he set the parfait as compared to the nude pictures. It is important to keep in mind that the prediction of the direction of relationship between the body image and T-T variables was such as to expect the trend in the Nonrepression group to be opposite to that in the Repression group.

5. Heart. Mean BFQ Heart scores were 5.9 ($\sigma = 4.1$) for the Repressors and 3.2 ($\sigma = 3.4$) for the Nonrepressors.

The T-T index chosen to tap the subject's reaction to a theme of openly displayed sexuality was the difference between the rank of his setting of the nude female picture and the rank of his setting of the clothed female picture. His response to a minimally sexualized clothed female figure could be compared with that to a nude maximally sexualized female figure. In the Repression category the mean T-T index was $-.4$ ($\sigma = 3.1$) and for the Nonrepressors it was $+.1$ ($\sigma = 2.7$). The chi-square describing the relation of BFQ Heart and the difference between nude female and clothed female ranks was significant ($\chi^2 = 4.9$, $df = 1$, $p < .05$). The higher the subject's Heart score the more likely he was to set the nude female larger (closer) than the clothed female. In the Nonrepression group the equivalent relationship was also significant ($\chi^2 = 6.0$, $df = 2$, $p = .05$); and, as anticipated, it was in the opposite direction. Contrasting with the trend for the Repressors, one finds here that the greater the Heart awareness the smaller (further away) the nude is set as compared to the clothed female.

Discussion. Four of 10 predictions regarding the relationships of the body attention and T-T picture setting variables were significantly supported, and 2 were supported at a borderline level. Within the Repression sample favorable results were found for four of the five hypotheses; while only two of five were favorable for the Nonrepressors. Apparently, the reaction aroused by the T-T task in the Re-

pressors results in a kind of involvement permitting body-image related attitudes to have an impact on their judgments of the pictures. Such involvement is much less evident for the Nonrepressors. The fact that there are differences in results revolving about the Repression-Nonrepression distinction confirms previous reports that the subject's type of involvement in the T-T task plays a role in the perceptual defense strategy he displays. A number of the findings, though significant, were really of a borderline character. However, what is impressive is the fact that the body attention variables predicted the direction of response trends for so many aspects of a small number of pictures. The results are encouraging as exploratory indications that the way in which an individual distributes his attention to his body is reflected in selective Thereness-Thatness perception.

The findings for BFQ Heart were the most consistent, with significant trends present for both the Repressors and Nonrepressors. In the former group the greater the individual's heart awareness the larger (closer) was his T-T setting of the nude female as compared to the clothed female, and in the latter group the obverse relationship appeared. This exaggerated response to the image of the nude female supports previously cited data which suggested that heart awareness is related to issues of puritanical morality with accompanying concern about sexual propriety.

Moderate support was obtained for the hypotheses related to BFQ Right and BFQ Back. In both instances there were significant results in the predicted direction in the Repression sample, although not so in the Nonrepression sample. The higher a Repressor's BFQ Right score the larger (closer) was his setting of the female nude in relation to the male nude. However, the higher his BFQ Back score the larger was his setting of the male nude relative to the female nude. Awareness of the right side of the body goes along with a defensive reaction to a heterosexual stimulus, and awareness of the back is accompanied by defensiveness in dealing with the "homosexual" image.

The results pertaining to Prominence and BFQ Eyes tended in the predicted direction, but not significantly so. The Prominence data were within a hairsbreadth of supporting the hypothesis in the Nonrepression category but not in the Repression category. The fact that the poorest results occurred for two variables which were both appraised with T-T response to the ice cream parfait (oral theme) raises the question whether this picture was inadequate in its arousal properties. The T-T pictures were achromatic drawings, and there are hints that it may be difficult to depict food vividly without the use of color. Analogous difficulty would not seem to apply to the same extent in representing heterosexual or homosexual themes. This matter will be explored further by making use of colored food pictures in future work.

Discussion of Overall Findings

Despite gaps in the data, there is a train of consistency and overlapping support for hypotheses which indicates that the manner in which an individual distributes attention to his body is related to his conflicts and defenses. Relationships have been demonstrated that diversely involve total body awareness, broad spatial dimensions of the body (viz., right-left, front-back), and specific organs (viz., eyes, heart). The allocation of body attention seems to be a carefully monitored process in which distinctions are made between sectors in terms of their meanings and valences. Indeed, the specificity of the conflicts associated with focusing upon certain body areas highlights the feasibility of using a person's body perceptions to obtain information about his personality. What are the origins of the relationships that apparently exist between allocation of body awareness and personality variables? What significance do these relationships have for the behavior of the individual? Currently one must approach such questions speculatively.

It is possible to conceive of multiple ways in which psychological attitudes and

conflicts might become linked with amount of attention devoted to a body area.

1. The link might derive from the fact that the same parental attitudes which shape one's personality early in life are also expressed in the treatment and restrictions applied to areas of one's body in terms of cleaning, touching, watching, "covering up," and so forth. For example, parental attitudes which declare the badness of sex might encourage traits like shyness or passivity in a man and also produce conditions which by preventing him from freely seeing and touching his genitals lead to vague awareness of, or decreased attention toward, them. This sort of "historical" explanation is, of course, particularly favored in the psychoanalytic literature.

2. Another possibility is that one's body is uniquely close to the ego or self and therefore likely (as is true of many ego-significant targets) to become in whole or part a "screen" upon which one projects attitudes about self and the world. This would be exemplified by the individual who feels inferior and therefore unrealistically perceives his body or some part of it as small.

3. A third alternative would liken awareness of a body area to the experience that goes along with tensing a part of one's body in preparation for an act. An angry person preparing to hit someone might perceive unusual tension in his biceps as he gets ready to swing. In this way, a persisting wish to obtain certain goals might be accompanied by a chronic increased subjective awareness of the "tensed" or "alerted" body areas used in attaining such goals. A relationship of this type between body awareness and psychological parameters can, of course, be conceptualized in terms of the immediate situation without appeal to early developmental influences.

4. Obversely, awareness of a body area might represent not preparation for an act but rather watchful inhibition to assure oneself that an act will not be committed. The person afraid of acting out an angry wish might monitor the body areas involved in expressing aggression to make sure they are not used for that purpose. This formulation can be extended to include ambivalence about a goal. In such a case, the body awareness might alternately represent the perception of preparation to act and then the determination to inhibit the action.

5. Still another possibility is that the awareness of a body area might provide a devious way of partially satisfying a "forbidden" wish. Conceivably, an individual who had learned fear of direct sexual expression might gain some gratification from the persistent sensations associated with an apparently "uncontrolled" awareness of his genitals, even if anxiety were prominently involved. Many of the phenomena of hypochondriasis could fit this category.

6. Wishes or fantasies which are forbidden might arouse guilt and the expectation that retaliatory damage will be applied to body areas involved in the satisfaction of the wishes. Body awareness in this instance would represent an anxious watching and guarding of an area to make sure that it did not get attacked or damaged. Freud's concept of "castration anxiety" illustrates this possibility.

7. Most speculative of all, as described in a previous paper (Fisher, 1959) is the view that an individual's awareness of a body area might reflect his incipient attempts to "try out" body responses or modes of expression related to a wish. A person with repressed anger who for some reason became less blocked in this respect might develop increased awareness of muscles in his arms as he privately (unconsciously) rehearsed what it would feel like to hit someone. Or a person who was guilty about incorporation might develop heightened awareness of his mouth as the result of greater freedom to "try out" the sensations of "taking in" and biting. The "trying out" could actually be a way of deciding whether to go on to more open and active forms of the response. There is some similarity between this concept and formulations in which thinking is equated with

minute muscular movements in the throat and elsewhere (e.g., Washburn, 1916).

If one considers the multiple ways in which body perception might become tied in with goals and wishes, it is clear that one is dealing with a system of high complexity. What general role might this system play in behavior? It has been suggested that there are specialized ways in which the awareness of a body region could complement a wish or conflict. For example, it might serve as a substitute for other kinds of body experiences or permit a covert "trying out" of new forms of wished-for expression. But in viewing the function of the system as a whole, it has been conjectured elsewhere (Fisher, 1965b) that it provides persisting signals that introduce selectivity in cognition and perception. The patterned awareness of certain body parts which have defined connotations over time may be regarded as an organized complex of peripheral cues which guide responses. A brief quote from a paper by Fisher (Fisher, 1965b, p. 539) summarizes this idea:

> The body scheme may be conceptualized as a representation in body experience terms of attitudes the individual has adopted. These are experiences coded as patterns of body activation (e.g., involving muscle, stomach). It may be presumed that the patterns of body activation exist as circuits based on the following sequence: perceptual focus upon a body area because of its utility or significance or activation in relation to a goal; increased physiological and also sensory arousal of the area as a consequence of its special prominence; further feedback from such arousal to the subsystem in the CNS involved in the original highlighting of the area. Thus, the individual's body scheme contains landmarks which reiterate to him that certain things are important and others are not. Just as a contracting stomach is a signal to seek food, the perceptual prominence of certain muscles maintained at high tonus may be a reminder to attend or not to attend to some class of objects.

This formulation grew out of a series of studies in which it was demonstrated that intensity of awareness of specific body areas was predictive of selective cognition and memory for certain classes of information.

In these terms, one may regard each of the body sectors in the present study which has shown a meaningful connection with personal attitudes as part of a system providing sensory cues which "feed back" to guide interpretation of the environment. The awareness of the back might serve to "remind" the individual that he must maintain self control, not "soil," and avoid certain types of relationships with men. The prominence of the heart might repeatedly signal the importance of behaving in a virtuous way and avoiding stimuli that are "bad" temptations. Or sensations from one's eyes could function to inhibit approach to situations which would stimulate oral incorporative wishes. The possibility that body sensations might function to modify perception and cognition in this fashion has been considered at length by Solley and Murphy (1960). They noted that Solomon and Wynne (1954) had found that avoidance conditioning in dogs was influenced by amount of visceral autonomic feedback. Solomon and Wynne had actually concluded that "at least some of the afferent feedback impulses from the viscera have the properties of stimuli and so are capable of becoming conditioned stimuli and drive stimuli [p. 369]." Solley and Murphy indicated that proprioceptive feedback might function similarly; and proposed that a percept could get linked or locked to proprioceptive and autonomic feedback mechanisms "so that the percept and the feedback mechanisms are mutually excitatory [p. 243]." Illustratively, it was suggested that the continuous tightening of certain muscles might act chronically to inhibit specific anxiety arousing memories. "The painful memories are 'locked' in a state of unawareness by the incessant feedback from the tightened muscles [p. 244]." Quite analogously, other investigators have found that cues derived from body sensations may influence perception and learning. Level of muscle tonus, asymmetry of tonus, position of body in space, body deformity, amount of autonomic arousal, and body sensations derived from drug effects have all proven to be variables that significantly affect cognitive processes (e.g., Belleville, 1964; Calloway & Dembo, 1958; Hinckley & Rethlingshaber, 1951; McFarland, 1958; Werner & Wapner, 1952). The body image

emerges in the model just described as a framework of meanings assigned to body areas which in turn are accompanied by sensory signals that have a selective impact upon perception and cognition. In assigning such importance to body signals outside of the CNS, one swings back in the direction of peripheral theories of thought and affect which were supported vigorously at one time by Titchener (1924), Washburn (1916), James (1892), Jacobson (1929), Freeman (1948), Guthrie (1952) and others. While theories emphasizing central factors are currently more acceptable and in vogue, it has been observed by several reviewers (e.g., Gellhorn, 1964) that the contribution of peripheral factors has not yet been adequately evaluated.

Summary

The major findings that have emerged are as follows:

1. The greater a man's focus of attention upon the right side of his body, the less active he is heterosexually as defined by his own self-reports; the higher is his level of anxiety about stimuli with threatening sex-role connotations; and the less free he is in expressing sexual ideas. It was not possible to show a consistent relationship between right awareness and anxiety about sexual themes as depicted by a measure derived from the Blacky Pictures.

2. Degree of attention directed by the subject to the back of his body proved to be positively correlated with the following: his motivation for exercising careful control in the expression of impulses; his tendency to convert hostility into negativism; his level of anxiety about anal themes; and his popularity among peers as measured by his self-reports. The intensity of his back focus was negatively correlated with his recall of how openly his father acted out anger, but the relationship was of a chance order with respect to his recall of his mother's expression of hostility. The hypothesis was not supported that back awareness is negatively correlated with open aggressiveness. In general, the results accorded with the view that the front-back dimension is meaningfully linked with Freud's "anal character" typology.

3. It was possible to demonstrate that back awareness is greater in paranoid than nonparanoid schizophrenics. This finding was seen as congruent with Freud's formulation concerning the importance of homosexual conflict in paranoia.

4. The hypothesis was supported that the greater the focus of attention on the back as compared to the front of the body the relatively higher will be the physiological activation of the first in relation to the second (defined by skin resistance).

5. Intensity of eye awareness was negatively correlated with interest in eating and the selective tendency to recall food words. It was positively linked with degree of anxiety aroused by an oral sadistic theme in the Blacky Pictures series. Relatedly, it was negatively correlated with how generous father was recalled as having been. A similar nonsignificant trend was found with reference to the recall of mother's generosity. The data certainly suggest that one's degree of eye awareness is directly connected with how much conflict one has about incorporation.

6. The prominence of the individual's own body in his perceptual field also seems to be tied in with incorporative difficulties. Body Prominence scores proved to be negatively correlated with interest in food and number of foods liked and positively so with amount of anxiety apparently aroused by the Blacky Oral Eroticism picture. However, they were not significantly correlated with reactions to the Blacky Oral Sadism picture or with recall of parental selfishness.

7. There was good substantiation that heart awareness is positively correlated with the individual's religiosity and his perception of how religious his parents had been. It was also found to be related to difficulty in recalling words with guilt connotations, but was not correlated with response to the Blacky Guilt Feelings picture. It was erratically related to a

number of measures involving sexual attitudes and behavior. Finally, it was noted to be negatively correlated with aesthetic interests, as measured by the Study of Values. The findings were considered to affirm the view that heightened concentration of attention upon one's heart is part of an orientation which involves religiosity, narrowed perspective on the world, and increased guilt.

8. The body-attention dimensions were, to an encouraging degree, able to predict selective perceptual responses to pictured themes presented in the Ames Thereness-Thatness apparatus.

9. The overall results indicated that the individual's manner of distributing attention to his body is intimately related to the traits, conflicts, and personality defenses characterizing him.

REFERENCES

ABRAHAM, K. The influence of oral eroticism on character formation. (1927). In *Selected papers of Karl Abraham.* London: Hogarth, 1927.

ALEXANDER, F. *Fundamentals of psychoanalysis.* New York: Norton, 1948.

ALLEN, M. K. Personality and cultural factors related to religious authoritarianism. Unpublished doctoral dissertation, Stanford University, 1955.

ARONSON, M. L. A study of the Freudian theory of paranoia by means of the Rorschach test. *Journal of Projective Techniques,* 1952, **16,** 397–412.

BARNES, C. A. A statistical study of the Freudian theory of levels of psychosexual development. *Genetic Psychology Monographs,* 1952, **45,** 105–175.

BELLEVILLE, R. E. Control of behavior by drug-produced internal stimuli. *Psychopharmacologia,* 1964, **5,** 95–105.

BELOFF, H. The structure and origin of the anal character. *Genetic Psychology Monographs,* 1957, **55,** 141–172.

BLUM, G. S. A study of the psychoanalytic theory of psychosexual development. *Genetic Psychology Monographs,* 1949, **39,** 3–99.

BUSS, A. H. *The psychology of aggression.* New York: Wiley, 1961.

BYRNE, D., GOLIGHTLY, C., & CAPALDI, E. J. Construction and validation of the Food Attitude Scale. *Journal of Consulting Psychology,* 1963, **27,** 215–222.

CALLOWAY, E., & DEMBO, D. Narrowed attention. A psychological phenomenon that accompanies a certain physiological change. *Archives of Neurology and Psychiatry,* 1958, **79,** 74–90.

COUCH, A., & KENNISTON, K. Yeasayers and naysayers: Agreeing response set as a personality variable. *Journal of Abnormal and Social Psychology,* 1960, **60,** 151–174.

DOIDGE, W. T., & HOLTZMAN, W. H. Implications of homosexuality among air force trainees. *Journal of Consulting Psychology,* 1960, **24,** 9–13.

EPSTEIN, L. The relationship of certain aspects of the body image to the perception of the upright. Unpublished dissertation, New York University, 1957.

FENICHEL, O. *The psychoanalytic theory of neurosis.* New York: Norton, 1945.

FERENCZI, S. Stimulation of the anal erotogenic zone as a precipitating factor in paranoia (1911). *Selected papers of Ferenczi.* New York: Basic Books, 1955.

FISHER, S. Body image and asymmetry of body reactivity. *Journal of Abnormal and Social Psychology,* 1958, **57,** 292–298.

FISHER, S. Extensions of theory concerning body image and body reactivity. *Psychosomatic Medicine,* 1959, **21,** 142–149.

FISHER, S. Head-body differentiations in body image and skin resistance level. *Journal of Abnormal and Social Psychology,* 1960, **60,** 283–285. (a)

FISHER, S. Right-left gradients in body image, body reactivity, and perception. *Genetic Psychology Monographs,* 1960, **61,** 197–228. (b)

FISHER, S. Body image and upper in relation to lower body sector reactivity. *Psychosomatic Medicine,* 1961, **23,** 400–402. (a)

FISHER, S. Front-back differentiations in body image and body reactivity. *Journal of General Psychology,* 1961, **64,** 373–379. (b)

FISHER, S. Developmental sex differences in right-left perceptual directionality. *Child Development,* 1962, **33,** 463–468.

FISHER, S. A further appraisal of the body boundary concept. *Journal of Consulting Psychology,* 1963, **27,** 62–74.

FISHER, S. Power orientation and concept of self height in men: preliminary note. *Perceptual and Motor Skills,* 1964, **18,** 732. (a)

FISHER, S. Sex differences in body perception. *Psychological Monographs,* 1964, **78,** 1–22. (b)

FISHER, S. Sex designations of right and left body sides and assumptions about male-female superiority. *Journal of Personality and Social Psychology,* 1965, **2,** 576–580. (a)

FISHER, S. The body image as a source of selective cognitive sets. *Journal of Personality,* 1965, **33,** 536–552. (b)

FISHER, S., & CLEVELAND, S. E. *Body image and personality.* Princeton: Van Nostrand, 1958.

FREEMAN, G. L. *The energetics of human behavior.* Ithaca: Cornell University Press, 1948.

FREUD, S. Character and anal eroticism (1908). *Collected Papers,* Vol. 2. London: Hogarth Press, 1924.

FREUD, S. Three contributions to the theory of sex (1910). In A. A. Brill (Ed.), *The basic writings of Sigmund Freud.* New York: Modern Library, 1938. Pp. 553–629.

FREUD, S. Psychoanalytic notes upon an autobiographical account of a case of paranoia (1911). *Collected Papers,* Vol. 3. London: Hogarth, 1950.

Gellhorn, E. Motion and emotion: The role of proprioception in the physiology and pathology of the emotions. *Psychological Review,* 1964, **71,** 457–472.

Guthrie, E. R. *The psychology of learning.* (Rev. ed.) New York: Harper, 1952.

Hastings, P. K. A relationship between visual perception and level of personal security. *Journal of Abnormal and Social Psychology,* 1952, **47,** 552–560.

Hastorf, A. H. The influence of suggestion on the relationship between stimulus size and perceived distance. *Journal of Psychology,* 1950, **29,** 195–217.

Hinckley, E. D., & Rethlingshafer, D. Value judgments of heights of men by college students. *Journal of Psychology,* 1951, **31,** 257–262.

Ittelson, W. H., & Kutash, S. B. (Eds.) *Perceptual changes in psychopathology.* New Brunswick: Rutgers University Press, 1961.

Jacobsen, E. *Progressive relaxation.* Chicago: University of Chicago Press, 1929.

James, W. *Psychology.* New York: Henry Holt, 1892.

Jasper, H. H., & Raney, E. T. The physiology of lateral cerebral dominance. *Psychological Bulletin,* 1937, **34,** 151–165.

Jourard, S. M., & Secord, P. E. Body-cathexis and the ideal female figure. *Journal of Abnormal and Social Psychology,* 1955, **50,** 243–246.

Kilpatrick, F. P. (Ed.) *Human behavior from the transactional point of view.* Hanover, N.H.: Institute for Associated Research, 1952.

Krout, M. H., & Tabin, J. K. Measuring personality in developmental terms: The Personal Preference Scale. *Genetic Psychology Monographs,* 1954, **50,** 289–335.

McFarland, J. H. The effect of asymmetrical muscular involvement on visual clarity. Paper presented at the Eastern Psychological Association Meeting, New York, 1958.

Meketon, B. W., Griffith, R. M., Taylor, V. H., & Wiedman, J. S. Rorschach homosexual signs in paranoid schizophrenics. *Journal of Abnormal and Social Psychology,* 1962, **65,** 280–284.

Miller, D. R., & Stine, M. W. The prediction of social acceptance by means of psychoanalytic concepts. *Journal of Personality,* 1951, **20,** 162–174.

Moore, R. A., & Seltzer, M. L. Male homosexuality, paranoia, and the schizophrenias. *American Journal of Psychiatry,* 1963, **119,** 743–747.

Murray, H. A. *Explorations in personality.* New York: Oxford University Press, 1938.

Popper, J. M. Motivational and social factors in children's perceptions of height. Unpublished doctoral dissertation, Stanford University, 1957.

Reich, W. *Character analysis.* New York: Orgone Institute Press, 1949.

Reiff, C. G. An investigation of relationships among body image and some ego functions involved in formal thought processes. Unpublished doctoral dissertation, New York University, 1962.

Ross, W. D. The Rorschach performance with neurocirculatory asthenia. *Psychosomatic Medicine,* 1945, **7,** 80–84.

Schilder, P. *The image and appearance of the human body.* London: Keegan, Paul Trench, Trubner, 1935.

Schneider, D. E. The image of the heart and the synergic principle in psychoanalysis (psychosynergy). *Psychoanalytic Review,* 1954, **41,** 197–215.

Schoen, Z. J., & Scofield, C. F. A study of the relative neuromuscular efficiency of the dominant and non-dominant eye in binocular vision. *Journal of General Psychology,* 1935, **12,** 56–181.

Secord, P. F. Objectification of word-association procedures by the use of homonyms: A measure of body cathexis. *Journal of Personality,* 1953, **21,** 479–495.

Shellow, R. S. Perceptual distortion in the spatial localization of emotionally meaningful stimuli. Unpublished doctoral dissertation, University of Michigan, 1956.

Solley, C. M., & Murphy, G. *Development of the perceptual world.* New York: Basic Books, 1960.

Solomon, R. L., & Wynne, L. C. Traumatic avoidance learning: The principles of anxiety conservation and partial irreversibility. *Psychological Review,* 1954, **61,** 353–383.

Starcke, A. The reversal of the libido-sign in delusions of persecution. *International Journal of Psychoanalysis,* 1920, **1,** 231–234.

Tausk, V. On the origin of the influencing machine in schizophrenia. *Psychoanalytic Quarterly,* 1933, **2,** 519–556.

Thurstone, L. L. *Examiner manual for the Thurstone Temperament Schedule.* Chicago: Science Research Associates, 1953.

Titchener, E. B. *The psychology of feeling and attention.* New York: MacMillan, 1924.

Travis, L. E., & Herren, R. Y. Studies in stuttering. V. A study of simultaneous antitropic movements of the hands of stutterers. *Archives of Neurology and Psychiatry,* 1929, **22,** 487–494.

Van Lennep, D. J. Projection and personality. In H. P. David & E. Von Bracken (Eds.), *Perspectives in personality theory.* New York: Basic Books, 1957. Pp. 259–277.

Von Ophuijsen, J. H. On the origin of the feeling of persecution. *International Journal of Psychoanalysis,* 1920, **1,** 235–239.

Wapner, S., & Krus, D. M. Behavioral effects of lysergic acid diethylamide (LSD-25). *A.M.A. Archives of General Psychiatry,* 1959, **1,** 417–419.

Washburn, M. F. *Movement and mental imagery.* Boston: Houghton Mifflin, 1916.

Werner, H., & Wapner, S. Toward a general theory of perception. *Psychological Review,* 1952, **59,** 324–338.

Wittkower, E., Rodger, T. F., Wilson, A. T. M. Effort syndrome. *Lancet,* 1941, **1,** 531–535.

(Received June 18, 1965)

Vol. 80, No. 10 Whole No. 618, 1966

Psychological Monographs: General and Applied

SIMULTANEOUS AND SUCCESSIVE CONTRAST EFFECTS OF REWARD MAGNITUDE IN SELECTIVE LEARNING[1]

NORMAN E. SPEAR AND JOSEPH H. SPITZNER

Rutgers University

4 experiments investigated the influence of a given magnitude of reward as a function of S's contemporary or previous experience with a different reward magnitude. The orthogonal variables studied included type of initial experience with reward (consummatory versus consummatory-plus-instrumental), response measure, apparatus, and intertrial interval. In addition to several points relevant to method, this research determined the following: (a) Latent learning of reward magnitude may be reflected as a simultaneous-contrast effect. (b) Reduction in reward associated with 1 stimulus may be accompanied by reduction in the rate of responding to another stimulus. (c) A Simultaneous-contrast effect exists when choice is between some and no reward. (d) A simultaneous-elation effect does not appear corresponding to the simultaneous-depression effect. (e) Behavior following a shift in the magnitude of reward associated with a given stimulus may be determined in part by the magnitude of reward previously associated with another, discriminably different, stimulus.

The effect on behavior of a given condition of reinforcement depends upon the subject's (*S*'s) history with other or the same conditions of reinforcement. One convincing example of this has been the demonstration of "contrast effects" (CEs) following shifts in the magnitude of reward. These have been labeled "depression" and "elation" effects by Crespi (1942) and are defined when the performance of *S*s exposed to such shifts drops below or rises above the performance of controls exposed to only a single reward magnitude.

Contrast effects generally have been interpreted from either of two theoretical frameworks. Crespi (1942), Bower (1961), and others have viewed them as emotional effects. Others (e.g., Bevan, 1963) have considered a perceptual source of CEs. These interpretations, however, have been hampered by a lack of consistent data and a narrow range of response measures. The present experiments were part of a series intended to add depth and breadth to the available relevant data. The strategy was similar to that of Spear (1964), Spear and Hill (1965), and Spear and Pavlik (1966); and the present experiments were directed specifically to a thorough examination of the paradigm for "simultaneous and successive contrast effects" (hereinafter referred to as SimCEs and SucCEs).

Spear and Hill (1965) considered two operationally distinct paradigms within the study of CEs of reinforcement conditions. In a SucCE, two successive stages of training differ only in terms of the conditions of reinforcement (in this case, magnitude of reward) associated with a particular stimulus-response (S-R) event. A SucCE is said to occur when performance during the second stage of training is inversely related to the magnitude of reward experienced during the initial stage. Tests of the SimCE, on the other hand, compare performance to two discriminably different stimuli which are associated with differential magnitudes of reward. In this paradigm, relative performance is measured when the stimuli are presented singly, and free choice is measured when the stimuli are presented simultaneously. A SimCE

[1] This research was supported by Grants MH-08888-01 and MH-08888-02 from the National Institutes of Health. M. P. Spear provided valuable advice and assistance in the writing of the manuscript.

is defined when performance to a given stimulus is inversely related to the reward magnitude associated with the alternative stimulus.

These paradigms were combined in a single design by Spear and Hill (1965). Rats experienced a large and a small reward in the respective alternatives of a T maze during the SimCE test. Then, for the SucCE portion, the larger reward was reduced. A SimCE was readily obtained, and the SucCE was numerically present in terms of running speed but not in choice behavior.

There are several advantages of combining the tests for SimCE and SucCE within one experiment. First, the relative robustness of these effects can be compared. The particular objectives of the Spear and Hill experiments did not permit an entirely adequate comparison in this respect. Their finding of a relatively weaker SucCE may have been due to the prior experience of *S*s with the Postshift reward magnitude and/or the accompanying SimCE. Indeed, either an emotion- or a perception-based explanation of CEs would predict this (see General Discussion). The present Experiment IV more adequately estimated the relative strengths of these CEs.

Combining the SimCE and SucCE paradigms also aids in determining the extent to which a CE is stimulus specific. Thus a tentative decision becomes feasible regarding the pervasiveness of the CE: Is a great deal of *S*'s contemporary behavior affected by a particular occurrence of a CE, or are relatively disjoint portions of his behavior unaffected? The former would be the case if, following a decrease in the reward on one alternative, it were found that *S*'s performance declined in terms of responses other than those directly associated with the shift (for example, if running speed were also found to decrease in the alternative in which reward is *un*changed). This possibility, though unsupported by the data of Spear and Hill, is important in view of the *response-ubiquitous* partial-reinforcement effect in extinction (Spear, 1964; Spear & Pavlik, 1966). This latter phenomenon is defined by increased resistance to extinction of a formerly continuously reinforced response as a consequence of *S*'s experience with some other partially reinforced event. Whether the CE is also ubiquitous throughout *S*'s behaviors can be determined by the combined test for SimCE and SucCE.

Finally, a combined test can measure CEs in terms of choice behavior. For example, when *S* is presented with two nominally equal alternative reward magnitudes, will he ever show *less* preference for the alternative which formerly was associated with a larger reward? The CE in lower animals has typically been estimated via a vigor measure, but inertial and physiological limits of such a measure have created methodological difficulties (cf. Knarr & Collier, 1962; Spence, 1956). A preference measure avoids these problems, and this act initially prompted the present use of the T maze.

Three general objectives covered in the following four experiments, then, are (*a*) to estimate the relative magnitude of the SimCE and SucCE, (*b*) to determine the extent to which relatively dissociated behaviors are affected by a reduction in reward, and (*c*) to further investigate the possibility of a SucCE in choice. Experiments I and IV concerned specific variables believed relevant to characteristics of the SimCE and SucCE, and Experiments II and III were conducted to clarify points of methodology.

EXPERIMENT I

The first experiment was concerned specifically with this point: Will a SucCE occur in choice behavior? The available data would require a negative answer (Spear, 1964; Spear & Hill, 1964, 1965). However, it seemed likely that the specific procedures employed may not have maximized the opportunity for this effect to appear. Of several problems, consider the following:

The occurrence of a CE in choices would appear to be a joint function of certain temporal properties of the SimCE and SucCE. Assume that choice is dictated by relative response strength in the alternatives and that response strength is some

linear function of running speed. Now, when S chooses between 12 and 1 pellets during Preshift and is shifted to 1 pellet on each alternative during Postshift, we know that two events will occur: response strength on the less favorable alternative (LFA) before the shift will be less than expected for 1 pellet, and response strength on the formerly more favorable alternative (MFA) after the shift will be about equally less than expected for 1 pellet. We also know that these response strengths will ultimately adjust to the same appropriate level. Thus, if the SimCE and SucCEs are equal in magnitude, the critical requirement for a CE in choices appears to be that the response strength to the LFA must adjust more rapidly than that to the MFA following the successive shift in reward. In other words, at least one of two events must occur: either the SucCE must depress MFA responding more than the SimCE depresses LFA responding, or S must recover from the SimCE sooner than from the SucCE. As a related consideration, note that the CE on choice would be defined when S responds with less than 50% preference for the originally MFA once reward is equated in the alternatives. For this to occur, the SucCE must either overcome the inertia of choosing the MFA (cf. Knarr & Collier, 1962) or outlast it. The probability of the latter is minimized by the transitive nature of the CE (see Gonzales, Gleitman, & Bitterman, 1962). Thus it would seem that the greater the likelihood of overcoming the inertia of choosing the MFA, the more favorable the conditions of a CE in choice behavior.

Therefore, the first experiment was performed within this general framework, and an attempt was made to minimize the SimCE while maximizing the SucCE. It was felt that this might be accomplished by giving S Preshift experience with the contrasting rewards associated with the distinctive stimuli but without having S respond differentially to the stimuli, except in terms of consummatory behavior. Thus, S would enter the Postshift stage without a history of differential running and choice behavior to the alternatives (including the typical SimCE) and without the strong initial tendency, or "response inertia," to turn toward the MFA from the choice point on free trials.

Therefore, in addition to the groups given conventional Preshift trials—Group 12-1 (receiving 12 pellets on one alternative and 1 on the other) and Group 1-1 (receiving 1 pellet on either side)—two comparable groups were only placed in, but not run to, the different goal boxes during Preshift. During Postshift, Ss in all groups were given conventional trials with only 1 pellet in each alternative. To the extent that the discrimination was established during Preshift training, a SucCE in choice behavior was expected to occur in the "placed" Ss since they were not subject to interference from initial response tendencies built up by choosing the MFA during Preshift.

Method

Subjects. The Ss were 64 experimentally naïve female albino rats of the Sprague-Dawley strain, approximately 60 days old at the start of training and weighing 180–200 gm.

Apparatus. The T maze, painted flat black except for the clear Plexiglass top, is shown in Figure 1. The 1-ft. start box was separated from a 1-ft. stem by a Plexiglas guillotine door. Each 2½-ft. arm contained a food cup 8 in. from its end. At the termination of each arm was a partition not quite as high as the top of the maze beyond which was located a dish of reward pellets to equalize any olfactory cues. Guillotine doors, used to prevent retracing, were located in each arm. One such door was 1 in. beyond the choice point, and the other was 21 in. from the choice point (1 ft. from the food cup). All interior sections were 4 in. wide and 4 in. high.

In order to make the alternatives more discriminably different, the floor and 1½ in. of the

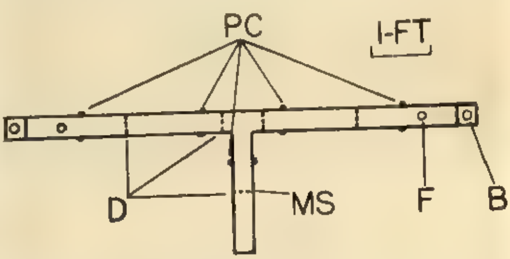

FIG. 1. T maze employed in Experiments I, II, and IV. PC = photocell; B = bowl of pellets included to mask odor; D = door; MS = microswitch; F = food cup.

walls of each arm from the food cup to the center of the choice point was covered with an interchangeable ¼-in. Masonite insert. The presence of inserts necessitated a ¼-in. "step up" as *S* entered the choice-point section. One set of inserts (always located in the right arm in Experiment 1) was built with the rough surface of the Masonite facing up and was painted dark brown. The second set of inserts (always in the left arm) had the smooth surface up and was painted white. Thus textural confounded with brightness (or color) differences were provided.

Times were recorded on Standard Electric Timers in .01 sec. Response measures included stem time (from the raising of the start door to the interruption of a photobeam 6 in. past the start door), turning time (from the first photobeam to the second located 6 in. beyond the choice point in either arm), committed time (from the second photobeam to the third located 24 in. past the second and 3 in. before the food cup), and choices on free trials. Times were converted to speeds by a reciprocal transformation, and speeds are reported in ft/sec.

The T maze was in a dimly lit cubicle which contained an exhaust fan providing masking noise as well as ventilation. The *Ss* were maintained, one to a cage, in a standard cage rack in the colony room under constant, bright illumination. When running the *Ss*, the experimenter (*E*) moved the cage rack into a dim passageway just outside the testing cubicle. On each trial, *S* was placed directly in the maze from its home cage and immediately returned at the completion.

Procedure. Upon arrival, *S* was placed on a deprivation schedule which was maintained throughout the experiment. The *S* received daily 10 gm. of finely ground Purina lab chow and 40 45-mg. regular Noyes pellets with ad lib access to water. Pellets consumed in the T maze during training were subtracted from the total, and the remainder, along with chow, was given in the home cage 20–30 min. after the daily trials. On Days 2–7, each *S* was prehandled (placed for 3 min. in a large, black box, during which time *S* was lifted and replaced by *E* five times).

The *Ss* were run in two replications of 32 *Ss* each, and each replication contained 8 *Ss* from each of the four groups. A 2 × 2 factorial design was employed, varying nature of Preshift trials and magnitude of reward on the MFA. During Preshift (the first 48 trials of training at 6 trials per day), Group 12-1-Run received 12 45-mg. pellets on the MFA and 1 pellet on the LFA. Group 1-1-Run received one pellet on either alternative during Preshift, the designations MFA and LFA being randomly assigned on the same basis as in 12-1.

Both Groups 12-1-Run and 1-1-Run received their Preshift trials in a typical manner, being placed in the start box and allowed to run to a goal box. Groups 12-1-Placed and 1-1-Placed received Preshift reward conditions identical to Groups 12-1-Run and 1-1-Run, respectively, but did not run to their rewards. Rather, on each trial, they were placed in the appropriate arm, facing the food cup about 3 in. from it. The *Ss* in all groups were removed from the maze as soon as the reward was consumed, provided they remained a minimum of 15 sec. and no longer than 3 min. The order of placements for Placed *Ss* was random during Preshift with the stipulation that the six trials of each day contain three placements to the MFA and three to the LFA. On each trial, Run *Ss* were put in the start box facing away from the start door. After 3 sec., *E* raised the start door regardless of *S*'s orientation. If *S* failed to stop a clock in 2 min., that and any other unstopped clocks were recorded as 2 min., and *S* was placed in the appropriate arm. If *S* made no choice on a free trial, he was randomly assigned to one arm. For the Run *Ss* during Preshift, Trials 1 and 5 of each day were free (*S* could enter either arm), and Trials 2 and 6 were forced (a closed door at the choice point prevented access to one arm) to the side opposite that chosen on the previous trial. Trials 3 and 4 of each day were forced randomly, one to each alternative, with the stipulation that half of the *Ss* in each group on each of Trials 3 and 4 be forced to the MFA and half to the LFA. Thus, equal experience to each alternative was ensured. During Postshift (Trials 49–120, Days 9–20), all *Ss* were run as were the Run *Ss* during Preshift, with the exception that one pellet reward was available on either alternative.

In assigning *Ss* to groups, the adjoining four cages in a cage rack contained one member from each group. These four *Ss* were run in rotation, resulting in a 3–4 min. intertrial interval. All such sets of four *Ss* were assigned the same MFA (left-White or right-Brown). Squads 1 and 2 in each replication were assigned left as the MFA as were Squads 7 and 8. Squads 3, 4, 5, and 6 of each replication were assigned right as the MFA. Thus one-half of the *Ss* in each of the four experimental groups had left-White as the MFA, and one-half had right-Brown as the MFA.

Results

Nature of specific analyses. Throughout this report, differences in choice behavior will be evaluated in terms of the stratified chi-square test, and running speeds (reciprocal times) by analysis of variance. The precise form of each analysis will be described only if it is not obvious. For example, all analyses of variance included replications and brightness of the MFA as sources, but this will not usually be noted. Replications never contributed a major source of variance to modify the effects cited, and so it will not be mentioned further. The reliable effects of brightness did not alter the conclusions in Experiment I and so they will not be considered there;

however, they will be discussed in detail in Experiment II.

Because of the large number of dependent variables, the Placed and Run groups are considered independently below. In addition, direct comparisons of the relative CEs following the Run versus the Placed condition are presented.

The Run groups represented a replication of Experiment II by Spear and Hill (1965), except for slight differences in the number of trials per day, the apparatus, and the employment of discriminably different alternatives. Accordingly, the results were virtually identical. The *Ss* in the present experiment ran uniformly slower on the average compared to the Spear and Hill results, but the relationships were the same. There initially was differential preference for the discriminably different alternatives, but this will be discussed in Experiment II.

Preshift. Choice probabilities, turning speeds, and committed speeds throughout

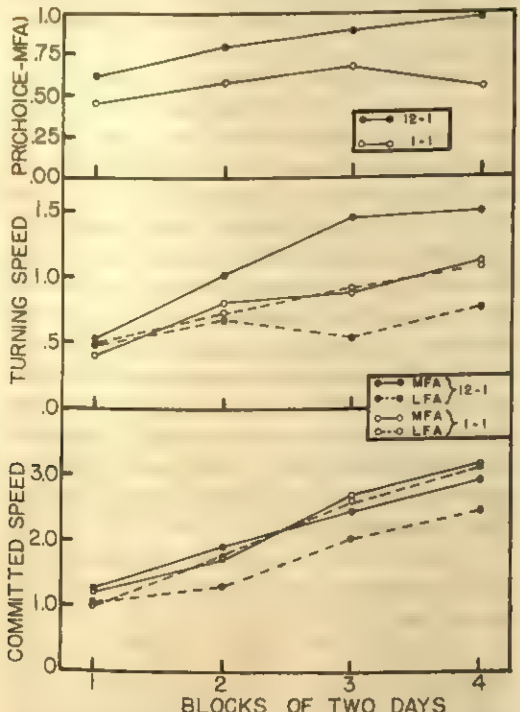

FIG. 2. Preshift choices, turning speed, and committed speed for "Run" *Ss* (i.e., *Ss* given conventional instrumental experience during Preshift) in Experiment I.

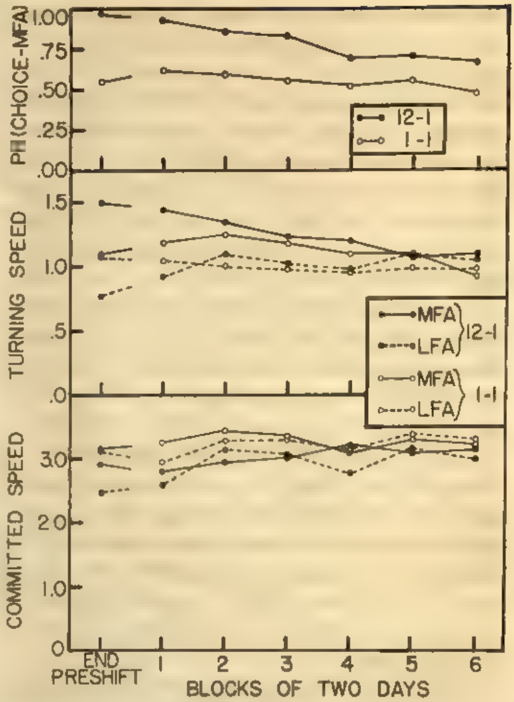

FIG. 3. Postshift choices, turning speed, and committed speed for "Run" *Ss* (i.e., *Ss* given conventional instrumental experience during Postshift) in Experiment I.

Preshift are shown in Figure 2. Preference for the MFA was, of course greater by Group 12-1 than by Group 1-1, $\chi^2(2) = 13.15$, $p < .005$.

The results in terms of turning and committed speeds were identical to those reported by Spear and Hill. On the last 4 days of Preshift, turning speed was greater to the MFA but less to the LFA for Group 12-1 relative to Group 1-1 ($p < .001$). The SimCE also occurred in committed speed—slower speed to the LFA was found by Group 12-1 compared to Group 1-1 ($p < .05$)—but committed speed to the MFA did not differ reliably between groups ($F < 1$).

Postshift: Run groups. Choice probabilities, turning speeds, and committed speeds are shown for the Run groups in Figure 3. It may be seen that the basic results obtained by Spear and Hill (1965) were also replicated in the Postshift stage of training by *Ss* in the Run groups. In particular, there was no tendency for the

12-1-Run group to show less preference than the 1-1-Run group for the MFA at any point. Also as in the Spear and Hill study, the turning speeds of the 12-1 group gradually adjusted to the appropriate level with no indication of a SucCE on the MFA. The committed speeds of Group 12-1 also eventually adjusted to the level of the baseline controls, although again there was some tendency for a SucCE on the MFA immediately subsequent to the shift in the reward magnitude there. In fact, mean committed speed over the first 4 days of the Postshift stage was less in Group 12-1 than in Group 1-1 on the MFA, $F(1,24) = 7.20$, $p < .02$; but this fact is difficult to interpret in view of the spuriously slow speed on the MFA by Group 12-1 at the end of Preshift. In any case, this may be added to the data accumulated by Spear and Hill (1964, 1965), which suggest a weak but continually appearing SucCE in this situation.

Postshift: Placed groups. The Postshift performance of the Placed groups shown in Figure 4 indicated that very little latent learning took place during the Preshift placements. Their speeds in the Postshift stage began at about the same rate as the speeds shown by the Run groups at the beginning of the Preshift stage. On the other hand, the choice data do suggest a slight preference for the formerly MFA in Group 12-1-Placed relative to Group 1-1-Placed during the first 4 days of Postshift, although this difference did not quite attain statistical reliability, $\chi^2(2) = 4.07$, $p < .15$.

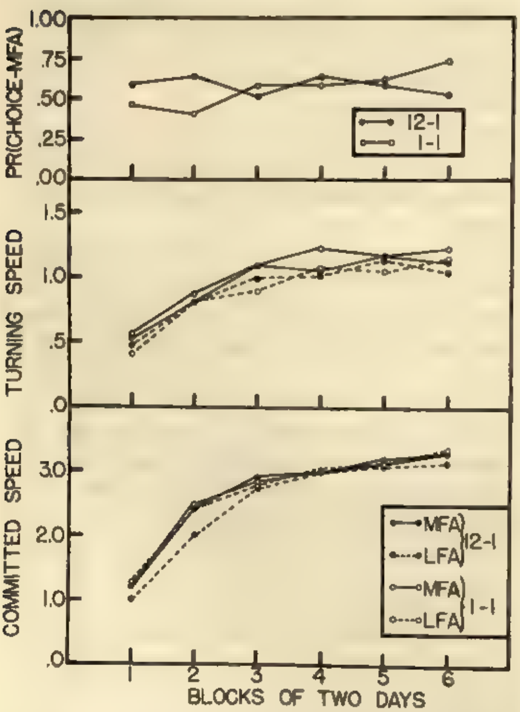

FIG. 4. Postshift choices, turning speed, and committed speed for "Placed" *Ss* (i.e., *Ss* given only consummatory experience during Preshift) in Experiment I.

It is clear in Figure 4 that the differential reward magnitude during the Preshift stage made little difference in terms of running speeds, with the exception of committed speeds to the LFA. It is notable that this SimCE, which was established during the Preshift placements, provided the only statistically reliable evidence that Preshift reward experience had an effect, shown by the fact that the speed to the LFA remained slower at this point for the 12-1 group relative to the 1-1 group, $F(1,24) = 6.11$, $p < .025$.

Postshift: Run versus placed groups. The combined results with the Placed and the Run conditions would seem to dictate one major conclusion: the tendency toward SucCEs in terms of running speed is more likely to occur following instrumental experience with the reward magnitude than following only consummatory experience. In fact, in every instance of a comparison in terms of running speeds, the SucCE was at least numerically greater for the Run conditions. This is consistent with data produced by Goodrich (1962), Goodrich and Zaretski (1962), and Spear (1965a) and with other unpublished runway data from our laboratory.

Consider the committed speed on the MFA during Days 1–4 of the Postshift stage (see Figure 5). The analysis revealed a statistically reliable SucCE overall, $F(1,56) = 4.77$, $p < .05$; and, as expected, the Run groups had greater mean speed overall than the Placed Groups, $F(1,56) = 127.66$, $p < .001$. The critical finding in this case was the interaction between re-

ward magnitude during Preshift and the Preshift treatment, $F(1,56) = 3.98$ ($F > 4.02$ is required for $p < .05$). In particular, the committed speed to the MFA was less for the 12-1 than the 1-1 groups in the Run conditions but was about equal in the Placed conditions. Thus the SucCE occurred in the Run conditions but not in the Placed conditions.

On the other hand, the SimCE, measured on the LFA as it carried over into Days 1-4 of the Postshift stage, did *not* vary between the Run and Placed conditions (see Figure 6). Overall, speeds to the LFA were less for the 12-1 groups than for the 1-1 groups at this point, $F(1,56) = 4.92$; $p < .05$, and speed was greater for the Run groups, $F(1,56) = 99.14$, $p < .001$. In contrast to the SucCE—and this is the critical point—there was no trace of an interaction between these conditions, $F(1,56) < 1$. That is, the carry-over of the SimCE into Postshift occurred about equally whether the *S*s had received instrumental or only consummatory Preshift experience. It is noteworthy that these Placed groups represent the only clear evidence that the SimCE may be greater than the SucCE.

In terms of choice behavior during the first 4 days of the Postshift stage, there were no statistically reliable differences in the effects of reward magnitude for the Run versus Placed conditions. In particular, there was no interaction between reward magnitude during Preshift and Run versus Placed conditions, $\chi^2(2) = .77$. This same analysis showed that the 12-1 groups preferred the formerly MFA with greater frequency than did the 1-1 groups, $\chi^2(2) = 10.75$, $p < .01$, and greater choice of the MFA was found in the Run groups, overall, than in the Placed groups, $\chi^2(2) = 10.75$, $p < .01$.

In absolute terms there was really no evidence for a SucCE in terms of choice behavior within either the Run or the Placed conditions. In no case did *S*s prefer the formerly LFA during Postshift. However, with the performance of Group 1-1 as the baseline rather than absolute preference, the relationship between the choice behavior of *S*s in the Placed and the Run conditions may be more reasonably compared. Recall that this experiment was originally designed with this comparison in mind. It was expected that more evidence for a SucCE in terms of choice behavior would be obtained in the Placed than in the Run conditions. The question, then, is whether the relationships between the 12-1 and 1-1 conditions differed at any point in Postshift for the Placed versus the

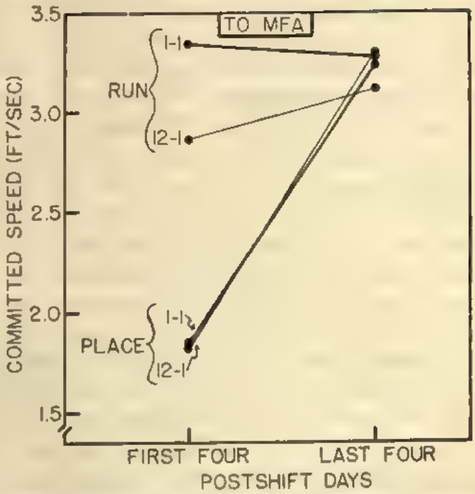

FIG. 5. The SucCE for *S*s placed in the differential goal boxes during Preshift compared to *S*s given conventional running experience in the maze during Preshift.

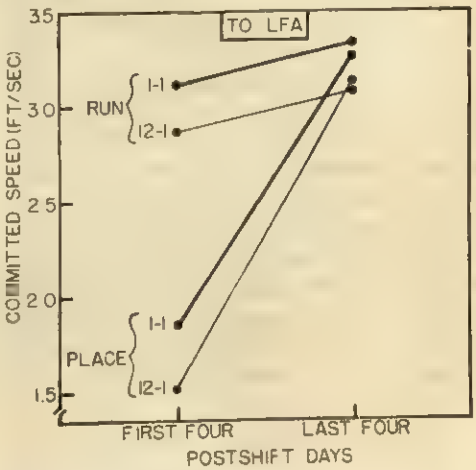

FIG. 6. The SimCE, measured in terms of its carry-over into Postshift, for *S*s placed in the differential goal boxes during Preshift compared to *S*s given conventional running experience in the maze during Preshift.

Run conditions. The answer would be "yes" if a significant interaction were obtained between reward magnitude and Preshift training conditions. Therefore, a chi-square test of number of correct choices was performed over the last 4 days of the Postshift stage. It was found that the predicted interaction did occur, $\chi^2(2) = 4.03$, $p < .05$, in the direction that the 12-1 condition chose the formerly MFA more often than the 1-1 condition within the Run groups, but the opposite was true within the Placed groups. However, this latter tendency toward a SucCE in terms of choice behavior within the Placed groups is spurious. As can be seen in Figure 4, this relationship within the Placed groups is not the result of less preference for the formerly MFA in Group 12-1 but rather it is a consequence of greater preference for the "dummy" LFA shown by the 1-1 group. Clearly not a great deal of weight can be placed on this finding, and it must be concluded that a SucCE in choices did not occur.

Discussion

Three major points of information were provided by this experiment. First, the SucCE in terms of choices was not obtained. Whether the *S*s had been placed or run during their Preshift experience did have some effect on the tendency toward a SucCE in choices, but the effect was marginal at best. Second, the SimCE again demonstrated its robustness by appearing equally whether the *S*s had had instrumental-plus-consummatory, or only consummatory, experience with the differential reward magnitudes prior to the reward shift. Third, although the SucCE in running speed was again mildly present after conventional Preshift experience, it was weakened and essentially erased when Preshift experience included only consummatory activity. Each of these three results is discussed below.

Contrast effects in choice behavior. Within the groups given conventional instrumental experience during the Preshift stage, there was no evidence that the 12-1 group ever preferred the formerly MFA with less frequency than did the 1-1 group. Had this occurred, it would have defined the SucCE in choice behavior. Including the two experiments reported by Spear and Hill (1965) and Experiments II and IV of the present report, there is now a total of five experiments in which the SucCE in choice behavior has not appeared following conventional Preshift training.

It is true that in the latter stages of the Postshift experience, the SucCE was numerically defined in the Placed groups but not in the Run groups, and this interaction was statistically reliable. However, the above-chance preference for the MFA by the 1-1 Placed group limited the implications of this fact. Moreover, it is also suspicious that this CE in choice behavior should not occur until so late in the Postshift stage, a point at which the CEs in speed have typically disappeared. Finally, it seemed strange that the "contrast effect" in choice behavior was not accompanied by the typically more sensitive CE in *running speed* within the Placed groups.

All things considered then, it was concluded that a CE in terms of choice behavior had not been demonstrated within this experimental paradigm. It may be possible to obtain such an effect on a position discrimination task by increasing drastically the number of Preshift trials (cf. Birch, 1964; Vogel, Mikulka, & Spear, 1966), or perhaps by using a visual discrimination task. These possibilities are currently being pursued but will not be considered in the remainder of this report.

Contrast effects in running speeds. The occurrence of the persisting SimCE subsequent to only consummatory experience during the Preshift stage—an instance of "latent learning" of reward magnitude—was important for several reasons. First, this represented the only known occurrence of a CE in running speeds subsequent to this limited kind of experience with the differential rewards. All the previous experiments that have attempted to show a CE subsequent to initial consummatory experience with the particular reinforcement (previously only the SucCE had been tested in this way) have failed to do so when a running-speed measure was employed (e.g., Goodrich, 1962; Spear, 1965a).

Of course, such CEs apparently are readily obtained with a bar-press response measure (Collier & Marx, 1959).

Second, this fact was especially important in relation to interpretations of CEs which require instrumental experience during the Preshift stage (cf. Pereboom, 1957). The finding of a CE after only consummatory experience with the rewards requires that theoretical emphasis be placed squarely on the stimulus properties of the reward itself, whether preingestive or postingestive.

Finally, it was important that the SimCEs occurred equally in the Placed and Run conditions, but that the SucCE did not. The weak trace of the SucCE shown in the Run conditions was completely absent in the Placed conditions. Although the implications are not entirely clear, this fact does suggest the possibility that the SucCE is governed by processes that are different from those responsible for the SimCE (see Discussion of Experiment IV).

One explanation of the lesser SucCE when only consummatory experience was given during the Preshift stage might emphasize the "memory," or in Capaldi's (1963) language, the "aftereffects," of the Preshift reward magnitude. It is clear that the occurrence of the SucCE is strongly dependent upon the retention of the aftereffect (or at least *some* representational response) of the Preshift reward. It may be that the rate of the running response itself during Preshift is an important component of this aftereffect. Thus, when only Preshift consummatory experience is given, this component is absent and the aftereffects of the Preshift reward are less available for comparison during the Postshift stage. It should be clear that the retention requirement is not as great in the SimCE as in the SucCE paradigm.

Although Experiment I demonstrated the effect of type of Preshift experience and replicated the basic phenomena obtained by Spear and Hill, there were three features of methodology that remained to be clarified. One of these features was the differential preference for the Brown versus the White alternative and its effect on CE phenomena: Does the effect of a shift in reward interact with *S*'s operant level of responding? The second was a troublesome feature that pops up from time to time in this kind of experiment, which we labeled the "tracking" phenomenon. The third, tested in Experiment III, concerned the possibility that the SimCE might be an artifact of the T maze.

EXPERIMENT II

A primary reason for this second experiment was to clarify two points regarding methodology. In doing so it was necessary to closely replicate the experimental conditions of the Run groups in Experiment I. First, certain potentially interacting effects concerning the differently appearing alternatives appeared interesting enough to warrant a closer look at these phenomena with an increased sample, thus providing a more powerful test. Second, in the experiments by Spear and Hill (1965), in Experiment I of the present studies, and in several other investigations from our laboratory, it had been noted that control groups not shifted in reward magnitude tended to behave as if they, too, had been shifted along with the experimental groups. They tended to behave as did the experimental rats immediately preceding them in the maze, and this behavior we labeled "tracking."

We had been aware of this latter possibility for some time and had taken measures to guard against the occurrence of *E* bias and systematic *E* errors. And, of course, the baseline control groups, such as Group 1-1, were always included instead of depending upon an absolute baseline. Still, there remained a needling tendency for the Group 1-1 controls to increase their probability of choosing the arbitrarily determined MFA when, for example, the majority of *S*s in the 12-1 group chose it. This was most apparent when every *S* in a given rotation was assigned the same side of the T maze as its MFA. Naturally, there was always an equal number of *S*s from each experimental condition represented within a given rotation, so the conditions would be equally affected. To the extent that all *S*s in a given rotation had a common MFA,

however, a sort of tracking phenomenon occurred. Tracking may have been exhibited in Experiment I of the Spear and Hill paper by the tendency for Group 1-1 to choose the arbitrarily designated MFA with a probability greater (numerically) than chance during the Preshift stage. It also may have appeared during the Postshift stage in Experiment II of that paper, as Group 1-1 decreased their choice of the arbitrarily designated MFA at about the same rate as Group 12-1.

Now it should be emphasized that if tracking did occur, it in no way confounded the conclusions, the reasonable assumption being that the effect of tracking is uniform over experimental conditions. All conditions were initially assigned their MFA in the same way, and baseline controls were always used.

The occurrence of this type of phenomenon is usually dismissed as a chance factor; indeed, the indications of tracking found in the Spear and Hill experiment did not attain statistical reliability. However, ethologists readily accept the possibility that rats may communicate via odors left in the maze and which may "help an animal to remember its way about" (Barnett, 1963, p. 31, p. 78). The intention of the present experiment was to maximize the possibility of such tracking behavior, but to restrict its source to odors more subtle than the occurrence of urine, feces, and mere number of rats that had preceded *S* down a particular path; thus, the present procedure included both the careful removal of urine and fecal traces, and the approximate equating of the number of previous rats that had gone to either alternative prior to a given *S*'s test. The critical difference was to be that half of the previous *S*s went to the LFA and half to the MFA. The question then was not just whether rats would follow the path of other rats, but whether they would follow them *differentially* to the MFA versus the LFA.

Method

Subjects and apparatus. The *S*s were 32 naïve female albino rats of the Sprague-Dawley strain, approximately 60 days old at the start of prehandling and weighing 180–200 gm. The apparatus and response measures were the same as those used in Experiment I, with the exception that there were two sets of colored inserts (discussed below).

Procedure. Maintenance, deprivation, and prehandling conditions were identical to those of Experiment I. The *S*s were fed a daily ration of 10 gm. of finely ground Purina lab chow, supplemented by 40 .045-gm. regular Noyes pellets (minus the number received in the T maze each day), in their home cages a minimum of 10 min. after the daily training. Water was always available.

The *S*s were randomly assigned to one of two groups and one of four rotations. As in Experiment I, Group 1-1 was run to 1 pellet on either alternative throughout the experiment, and Group 12-1 was run to 12 pellets on the MFA and 1 pellet on the LFA during Preshift, and to 1 pellet on either alternative during Postshift. Each rotation was defined in terms of the color and location of the assigned MFA and included four *S*s from each group. For two of the rotations assigned Brown as the MFA, one set of colored, T-maze arm inserts was employed; for one of these rotations, Brown was on the left, and for the other, it was on the right. For the other two rotations, White was assigned as the MFA, and the second set of inserts was employed; for one of these rotations, White was on the left, and for the other, on the right. Since it was hypothesized that rats leave differential spoors in accord with the "attractiveness" of the alternative, it was felt that the strength of these traces might summate across trials; that is, as more members of Group 12-1 ran to a given MFA and LFA, the differential spoors left behind by these *S*s would accumulate.

As in Experiment I, all *S*s were given six trials per day throughout the experiment. Trials 1 and 5 of each day were free trials, and Trials 2 and 6 were forced to the side opposite that chosen on the preceding trial. Trials 3 and 4 were forced, one to either alternative, with the stipulation that within each rotation and on any trial, half of the *S*s in each group were forced to one alternative and half to the other.

In order to detect tracking in Group 1-1 *S*s, a systematic running procedure was followed throughout training: each rotation was run as a unit, and several minutes separated the running of successive rotations. On each day, the four members of Group 12-1 in each rotation were given their first two trials before any members of Group 1-1 were run. Thus, each of the Group 12-1 *S*s ran one trial to each alternative before Group 1-1 *S*s entered the apparatus. This presumably maximized the presence of differential spoors, while equating the number of rats which traversed each alternative. For the next four trials, all eight *S*s in the rotation were run in succession as follows: first there were two *S*s run from Group 1-1, then four *S*s from Group 12-1, then two *S*s from Group 1-1. The same *S*s were run in the same order throughout training. Following these trials, the four *S*s in Group 1-1 were given their final

two trials in rotation. It was believed that the final two *S*s in Group 1-1 would show greater evidence of tracking than the first two, since on successive trials, they would immediately follow the Group 12-1 *S*s. The first two *S*s from Group 1-1 would, on trials after the first, receive their trials following the final two *S*s in Group 1-1. This procedure also insured that preceding a trial given an *S* from Group 1-1, an approximately equal number of rats had experienced either alternative. In addition, urine and feces were removed from the maze after each trial. Thus, cues based on differential *number* of previous *S*s and upon differential urine and feces (which could indicate emotional responses) were not available to *S*s from Group 1-1.

As in Experiment I, *S*s were run for 8 days of Preshift and 12 days of Postshift. It should be emphasized that the present design did not maximize the possibility of obtaining evidence for tracking behavior. The design was restricted by the additional aim of replicating and clarifying the result obtained in Experiment I relevant to the effects of color of alternative. Had the hypothesized "tracking" phenomenon been of major interest, a more sensitive test of its existence might surely have been devised.

Results

No convincing evidence for tracking, as defined here, was found. Several analyses were employed as tests of this phenomenon (within the limited design employed), and they all yielded negative results. Only a few of these analyses need be noted here as examples of the kinds of tests possible from this experiment.

First there were direct tests in which the absolute performance of *S*s in Group 1-1 could be inspected in terms of choice of the MFA (i.e., that alternative of the maze which was the MFA for the *S*s in Group 12-1 run in the same rotation as the respective *S*s in Group 1-1 relative to the LFA). It was found that *S*s in Group 1-1 chose both alternatives with about equal frequency during the Preshift stage and thus showed no tendency toward the 12-1 MFA. Moreover, during the Postshift stage there appeared no tendency for *S*s in the 1-1 group to decrease their frequency of choosing the formerly MFA in accord with the behavior of the *S*s in Group 12-1 (as apparently had been the case in the Experiment II of the Spear-Hill paper). In terms of running speed, Group 1-1 yielded no tendency during the Preshift stage toward faster turning or faster committed speed in the MFA than in the LFA. There occurred a slight tendency toward a decline in MFA turning speed during Postshift which was somewhat suggestive of a tracking effect in Group 1-1, but the lack of any similar effect in the committed speed made this result quite spurious. In terms of the relative effect of tracking on those Group 1-1 *S*s run before the Group 12-1 *S*s, compared with those Group 1-1 *S*s run after the Group 12-1 *S*s in a given rotation, there were again no reliable signs that tracking was contributing variance.

Second, recall the expectation that to the extent that tracking did occur, it would most strongly influence those *S*s in Group 1-1 which were run in the rotation immediately following the four *S*s in Group 12-1. None of the several analyses suggested this occurrence.

One other set of tests evaluated performance of Group 1-1 *S*s, expecting that any influence of tracking should have been more apparent during the later trials of the daily session than during the initial trials. These analyses also yielded no positive evidence for tracking.

Thus it is concluded that when the total number of rats running in either alternative is approximately equated and when urine and feces are removed from the maze, the performance of a given *S* is not seriously affected when run in the same rotation as other *S*s which have a common MFA and LFA.

Replication of past results. The choice behavior and running speeds measured in this experiment agreed with those obtained in previous experiments on SimCEs and SucCEs. The statistical analyses confirmed this replication in terms of all essential facts. The absolute values agreed nearly completely with those of Experiment I with the exception of a slight, though uniform, increase in MFA speeds by Group 12-1 of Experiment II. Therefore, there is no need to repeat the statistical particulars here.

The effect of brightness of alternative. It is a fact that in an apparatus such as was used in Experiments I, II, and IV of this report, albino rats have an initial (operant) preference for the darker of the alternatives. Is the effect of a reduction in

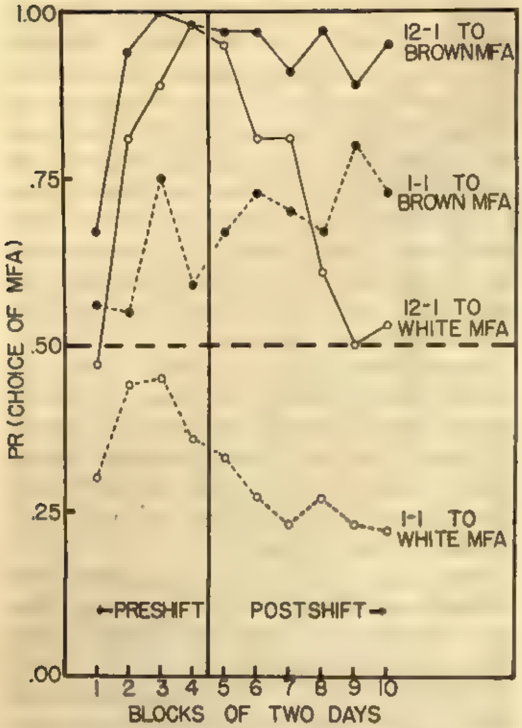

FIG. 7. Probability of choosing the MFA during Preshift and Postshift when the MFA was Brown compared to when it was White. Data from Experiment I (Run *Ss*) and Experiment II are combined—each point represents the mean for 32 *Ss*.

reward the same regardless of *S*'s operant level of performance? By assigning alternatives equally as MFA and LFA and combining data from comparable groups in Experiments I and II, a large sample could be used to provide an answer. Specifically, 32 *Ss* from each of Groups 12-1 and 1-1 were considered; half had been randomly assigned White as MFA and half Brown.

The particular question concerned the relative change in behavior in an alternative, concomitant with a change in conditions of reinforcement, when that alternative was initially the preferred, compared with when it was initially the unpreferred, alternative. For example, would *S* more rapidly decrease its response rate and/or preference in an initially *un*preferred alternative when the magnitude of reward is reduced in the alternative in question?

Preference for the Brown alternative. The first trial in the T maze was a free trial, and 61% of all *Ss* chose Brown on this trial. This preference did not decrease with succeeding experiences in the two alternatives, even when both alternatives were rewarded equally. In fact, preference for the Brown increased with training. During the last 2 days of the Preshift stage, *Ss* from Group 1-1 were choosing Brown on 67% of the free trials. Throughout the Postshift stage, 27 of the 32 *Ss* from Group 1-1 showed an overall preference for the Brown alternative. Averaging across Postshift trials, it was found that *Ss* from Group 1-1 chose the Brown alternative 73% of the time. These choice data may be seen in Figure 7.

Effect of brightness of alternative on conclusions concerning changes in behavior. When reward magnitude was reduced in the MFA, choice of the MFA by *Ss* in Group 12-1 was more greatly reduced when that alternative was White than when it was Brown (see Figure 7). This was confirmed statistically by a Mann-Whitney *U* test on the differences between the number of MFA choices by Group 12-1 during the last half of Preshift and the number of formerly MFA choices by this group during the first half of Postshift, $U = 38$, $p < .001$. Because Group 1-1 had been included to provide the appropriate baseline, it may be seen that the greater change in choice behavior when White was the MFA is simply a consequence of the fact that these *Ss* required a greater behavioral change to adjust to their baseline control (100% to 25%) than did *Ss* with Brown as the MFA (from 100% to 75%). Of course, we may not be dealing with an equal-interval scale here and probably are not. Nevertheless, it does appear that if one were to extrapolate the curves in Figure 7, both subgroups of Condition 12-1 would have adjusted to the level of the baseline control at about the same point in the hypothetically extended Postshift stage.

These facts are important in view of the erroneous conclusions that could result if the appropriate control groups had not been included. For example, if the baseline had not been established by Group 1-1, the more rapid change in choice behavior by Group 12-1 *Ss* with White as the MFA

might have taken on quite a different meaning. Or, had only Group 12-1 been included with Brown as the MFA, one might have concluded that choice behavior is relatively impervious to change as a consequence of a reduction in reward in one alternative so long as reward remains in both alternatives.

The question of relative change in behavior as a function of color of alternative may also be asked in terms of running speed measures. In certain respects, running speed is more sensitive to changes in reward magnitude than is choice behavior in this situation (Spear & Hill, 1965).

Figure 8 shows the course of running speed during Preshift for *Ss* having White as the MFA and for those having Brown as the MFA. Without going into detail, it may be said that no reliable interactions occurred between brightness of the MFA and reward magnitude. This was true in terms of both the rate of increase in speed during the first few days of Preshift and in terms of asymptotic speeds at the end of Preshift.

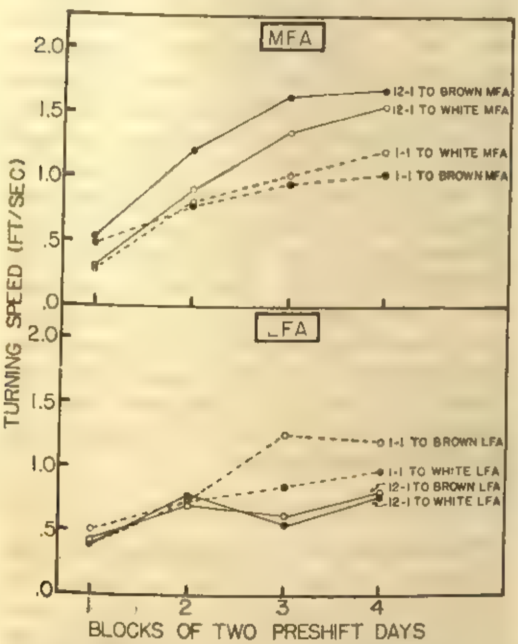

FIG. 8. Turning speed during Preshift when the MFA was Brown and the LFA White, and vice versa. Data from Experiment I (Run *Ss*) and Experiment II are combined—each point represents the mean for 32 *Ss*.

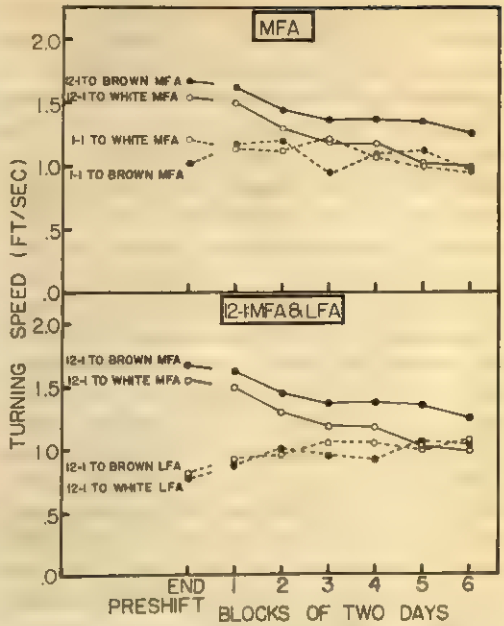

FIG. 9. Turning speed during Postshift combining the data from Experiment I (Run *Ss*) and Experiment II—each point represents the mean for 32 *Ss*. The top half shows MFA speeds for Group 12-1 with Brown MFA and Group 12-1 with White MFA in relation to their corresponding baseline control from Group 1-1. The bottom half shows the MFA and LFA turning speed for Group 12-1 when Brown, compared to White, was the MFA; the baseline here are LFA speeds by Groups 12-1.

Brightness of the MFA and LFA did, however, determine the adjustment of running speeds subsequent to the reduction in MFA reward from 12 pellets to 1 pellet. Turning speeds during the Postshift stage are shown in Figure 9 as a function of these variables. The top half of the figure shows the decrease in MFA turning speed by the 12-1 groups in relation to the baseline: here the baseline is defined in terms of the 1-1 control groups. The lower half of the figure again shows the decreasing MFA speeds by Group 12-1, but in this case the baseline is the corresponding LFA speeds of this group.

In view of the control groups and counterbalancing employed, it may be seen that the interacting effects of brightness with reward magnitude are not really serious for our purposes. However, it is clear in Figure 9 that the precise conclusions derived from an experiment of this kind

could be greatly influenced by the brightness of the MFA—or any similar factor creating differential operant preference—particularly if it were not balanced across conditions. For example, had only the White alternative been employed as the MFA in the top figure, one would have concluded that adjustment in turning speeds is completed rather rapidly—by the fifth and sixth days of the Postshift stage. On the other hand, had only the Brown alternative been employed as the MFA, the conclusion would have included some doubt as to whether speeds on the MFA would ever adjust to the baseline, since they gave no indication of doing so even after 72 Postshift trials. In terms of the final level of adjustment of turning speed to the MFA (last half of the Postshift stage), brightness of alternative did not reliably interact with Preshift reward, $F(1,60) = 1.69, p > .10$.

The bottom of Figure 9 illustrates the same principle but with the added complication that the baseline is defined in terms of the performance on the LFA by *S*s in Group 12-1. Since running speed to the LFA increased during the Postshift stage as a consequence of its recovery from the simultaneous depression effect, it clearly provides an inappropriate baseline for the definition of the SucCE. In this case, conclusions based upon turning speed during the last half of Postshift are clearly biased by brightness of MFA. A mixed analysis of variance (Brightness × Preshift reward) revealed a statistically significant interaction, $F(1,30) = 6.02$, $p < .025$, which reflected the convergence of MFA and LFA speeds for *S*s with the White MFA, in contrast to the continued separation of these speeds when the former MFA was Brown. These same general changes in behavior were also obtained in terms of the committed speeds, occurring more rapidly than those of turning speed.

Discussion

Experiment II yielded information relevant to: (*a*) the influence of tracking on behavior in the present T-maze situation; (*b*) replication of facts concerning CEs in a position discrimination, and (*c*) the necessity for including baseline control groups and for counterbalancing T-maze-alternative characteristics among MFA versus LFA assignment in this kind of research. These factors are discussed briefly below.

a. It was found that if such tracking does exist in the present situation, it is not an important source of variance. No evidence for tracking could be obtained when care was taken to remove obvious signs (feces, urine) from the maze and when a trial for a given rat had been preceded by an approximately equal number of trips to each alternative by previous rats in the maze.

b. The major facts obtained previously in experiments on the SimCEs and SucCEs were replicated. These included the findings of a reliable SimCE in running speed, a numerical but weak SucCE in running speeds, no SucCE in choices, and eventual adjustment to the level of the baseline control, in terms of all response measures, following the reward shift.

c. It was shown that the lack of independent control groups necessary to establish a baseline and the failure to balance out the initially preferable alternative when assigning the MFA and LFA could result in inappropriate conclusions concerning adjustment of behavior following a shift in the conditions of reinforcement. In particular, it was shown that if the initially less preferable alternative is the MFA and the more preferable alternative is the LFA, behaviors in these alternatives converge much more rapidly once the alternative reinforcement conditions are equated than when assignment of MFA and LFA is reversed. In the latter situation, there is relatively little change in behavior even after 72 Postshift trials.

EXPERIMENT III

This experiment, as Experiment II, was concerned primarily with problems of method and interpretation. There were still three points regarding the SimCE and SucCE paradigm that needed clarification. These points required an experiment which tested for SimCEs and SucCEs of reward magnitude, but with alternative stimuli

which were relatively independent. When the T maze was used, experience with one alternative excluded the possibility of experiencing the other. This was particularly true on free trials, and it could conceivably influence behavior on forced trials. In Experiment III, this was changed by using two distinctive runways (one White and one Black). Thus, only one possible "alternative" existed on each trial. Each rat experienced both runways. During Preshift, one runway was associated with the larger reward, the other always with the smaller reward; and both runways were associated with the smaller reward during Postshift.

The first point of interest was whether a SimCE could be measured during Preshift. Would *S*s run slower for the small reward in one runway when the larger reward was presented in the other runway than if the same small reward was available in both runways? If the SimCE did not exist in such a situation, one could always argue that its occurrence in the T maze was an artifact created on forced trials to the LFA in the 12-1 groups. Perhaps the attraction to the alternative MFA acted to "pull *S* back" from the LFA rather than to slow his progress to the LFA per se as is implied by the term "contrast effect." Other investigators have conducted studies corresponding to the Preshift stage of such an experiment. Goldstein and Spence (1963) found no evidence for such a phenomenon, but Bower (1961) and Bower and Trapold (1959) did. Because of this disagreement, a test employing our conditions seemed desirable.

The second point concerned the SucCE. Recall that the SucCE as measured in the present paradigm is weak and may lack statistical reliability within a given experiment. Perhaps some aspect of the T-maze apparatus was limiting its effectiveness. This possibility, therefore, was investigated by including the SucCE paradigm using the present dual runway apparatus.

Finally, the interacting effect of brightness of alternative was considered. It was possible that when the Black and White alternatives were experienced separately in the double runway situation, the effect of this variable might be relatively slight in comparison with the T maze in which *S* may compare Black and White simultaneously. Although Experiment III did not provide a rigorous test of this possibility, a rough estimate of the influence of Black versus White as an absolute source of variance was obtained.

Method

Subjects. The *S*s were 28 female albino rats of the Sprague-Dawley strain approximately 65 days old at the start of experimental training. All *S*s had been run by a different *E* in the same two runways prior to the present training under similar conditions of reinforcement with the exception that *all S*s had been differentially reinforced, receiving 10 pellets on the MFA and no reward on the LFA, following an initial series of rewarded trials in both alleys and a series of nonrewarded trials in both alleys. Approximately 20 trials had been given each *S* prior to the present training, which began 2 days after the last trial under the prior conditions. In the present study, all *S*s were assigned the same MFA as in the initial training.

Apparatus. The testing apparatus consisted of two Hunter runways which, briefly, consisted of a start box 5 in. wide and 1 ft. long; an alley 4 in. wide and 33 in. long; and a goal box 5 in. wide and 1 ft. long. All sections were 4 in. high and constructed of Masonite, except for the top and sides of the alleys and goal boxes which were Plexiglas. Raising the guillotine start door, which separated the start box from the alley, started a Standard Electric Timer which was stopped by the interception of a photobeam located 4 in. beyond the start door. Interruption of the first beam started a second clock which was stopped by the interruption of a second beam located 7 in. inside the goal box door separating the goal box from the alley and 3 in. before the food cup which was located at the rear of the goal box. One of the runways, designated White, was painted white throughout, except for the top, which was clear. The other runway, designated Black, was painted black throughout, except for a ½-in. strip along the top of the alley and goal-box portions of the runway.

Procedure. The *S*s were maintained on ad lib water, 30 45-mg. regular Noyes pellets, and 10 gm. of finely ground Purina lab chow daily. The pellets (minus those received in the runway) and chow were combined and given to *S* 15–20 min. after the daily training.

All *S*s were given four trials per day. The runway used on a particular trial was randomly predetermined with the stipulation that half of the trials of each day (two) be in each runway. The *S*s were run in rotations of four, resulting in approximately a 4-min. intertrial interval. Each *S* was removed and returned to its home cage upon consuming the reward, provided it remained a minimum of 15 sec. and no longer than 3 min.

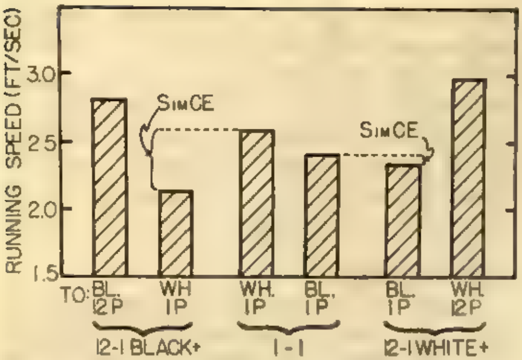

FIG. 10. Mean running speed during the last two days of Preshift for each group in Experiment III.

The *Ss* were equally balanced among groups on the basis of their experimental histories: 20 *Ss* were assigned to Group 12-1 and 8 *Ss* to Group 1-1. In addition, half of the *Ss* in each group were assigned Black as the runway with the larger reward (MFA); half were assigned White as the MFA. In the case of *Ss* in Group 1-1, the designation was arbitrary; these *Ss* received one 45-mg. pellet in either runway for the entire experiment. The *Ss* in Group 12-1 received 12 pellets in the MFA and 1 pellet in the LFA during Preshift, which comprised the first 11 days (44 trials) of training. During Postshift (Days 12–17; Trials 45–68), all *Ss* received one pellet reward on all trials in either runway.

Results and Discussion

In general, the SimCEs and SucCEs occurred with this present procedure just as they had when the T-maze apparatus was used. However, the effect of brightness was quite different in this procedure in which *S* was exposed to only one brightness at a time. Terminal running speed for the three conditions during Preshift may be seen in Figure 10. Speed to the MFA on the last 2 days of the Preshift stage was greater the larger the reward, both when the Black runway held the larger reward ($t = 2.13$, $p < .05$) and when White was associated with the larger reward ($t = 5.54$, $p < .01$). The SimCE occurred during these last 2 days of the Preshift stage as Group 12-1 had slower mean speed to the White LFA than Group 1-1 ($t = 2.29$, $p < .05$), although the numerical SimCE obtained by *Ss* in Group 12-1, which received their one pellet in the Black runway, did not obtain statistical significance at the .05 level. This is believed to be largely a consequence of the fact that this group had faster mean running speeds to begin with, as evidenced by running on the first few trials of Preshift.

Following the decrease in reward on the MFA for the 12-1 groups, behavior was consistent with that previously obtained in the T maze. In particular, running speed to the MFA decreased appropriately, while running speeds to the LFA were not so greatly affected. The successive depression effect was weak and its reliability not too convincing. In terms of total Postshift speed to the MFA, the performance of Group 12-1 with the Black MFA did undershoot that of the baseline control ($t = 2.10$, $p < .05$), but the performance of Group 12-1 with the White MFA did not differ from its baseline ($t = .80$).

Thus it was apparent that the SucCE was no stronger under these conditions than in the previous experiments with the T maze. Where it was reliably measured in Group 12-1 with the Black runway as the MFA, the CE was perhaps inflated by the spurious increase in running speed there by Group 1-1. Finally, it was clear that the decrease in behavior during the Postshift stage obtained in the White runway was no greater than that obtained in the Black runway; indeed, the opposite trend occurred.

The SimCEs and SucCEs in running speed, then, occurred in about the same way in the dual-runway situation as in the T-maze experiments. However, the effect of brightness when presented separately did not produce the same trend occurring in the T maze. In any case, not a great deal can be made of the relative interacting effects in the two situations: the dimension existing in the T maze was only approximated in the dual-runway situation, and the interacting effects in the T maze contributed only a minor source to the variance, anyway.

EXPERIMENT IV

The final experiment in this series was designed to answer certain general questions relevant to understanding the effects of a shift in magnitude of reward, and other questions arising from previous re-

sults with SimCEs. The following were considered: (*a*) To what extent is a SucCE modified as a consequence of *S*'s prior experience with a SimCE and/or the Postshift magnitude of reward? (*b*) Does a SimCE exist when a discrimination is formed between stimuli associated with *some* reward and stimuli associated with *no* nominal reward—the typical discrimination task? (*c*) To what extent does a simultaneous *elation* effect occur, and is it comparable in magnitude to the simultaneous *depression* effect? (*d*) What is the effect of distribution of trials on the simultaneous and successive depression effect when more than one trial per day is given? These four basic problems of Experiment IV are elaborated below.

a. The SucCE as a function of reinforcement history. The question was whether the extent of a SucCE is affected by previous experience with a SimCE and/or the magnitude of reward presented in the LFA during the Postshift stage. In this experiment, as before, Postshift performance to one pellet on either side of a T maze was compared as a function of the reward previously obtained in these alternatives during Preshift. Performance on the MFA as a function of prior LFA reward was of prime relevance. Specifically, this question was concerned with the relative successive depression effects obtained in the MFA for *S*s which had alternative Preshift rewards of 12 pellets and 12 pellets (Group 12-12), 12 pellets and 0 pellets (Group 12-0), and 12 pellets and 1 pellet (Group 12-1). From past data, it was certain that Group 12-1 would reveal a simultaneous depression effect during the Preshift stage. Group 12-1 also had Preshift experience with the Postshift reward of one pellet. Obviously, Group 12-12 would have neither of these experiences during Preshift. It was an empirical question whether *S*s in Group 12-0 would reflect a SimCE during the Preshift stage; certainly they would not have experienced the Postshift reward prior to the shift. Thus, this design permitted a comparison of the extent to which the SucCE would occur for *S*s which previously had experienced both a SimCE and the Postshift magnitude of reward, relative to *S*s who had experienced neither of these and to *S*s which had experienced only the SimCE.

Most critical concern was given this latter group—Group 12-0. If their performance reflected a SimCE during the Preshift stage, it would have suggested that these *S*s responded to zero reward as *if* it were, in fact, a small nominal reward. The alternative would be that response to zero reward is unique, that it reflects essentially zero behavior which is unaffected by *S*'s experience with other rewards elsewhere, and that it therefore should not enter into a SimCE. It is clear that if it were found that *S*s in Group 12-0 were responding to their LFA as if small reward were present there, their behavior should then have been more similar to that of *S*s in Group 12-1 than Group 12-12. On the other hand, if it were the case that the reinforcement conditions in the MFA and LFA operate independently on *S*s in Group 12-0—that is, if no SimCE exists—then there would have been no reason to expect the Postshift performance in the MFA by this group to be any different from that found in Group 12-12. It may be noted that these conditional predictions are essentially atheoretical, at least to the extent that both a perceptual and emotional interpretation of CEs would appear to make essentially the same predictions (see Discussion).

b. The SimCE on zero reward. The conventional discrimination task includes a choice between stimuli associated with some reward and stimuli associated with no reward. Does a SimCE exist under these circumstances? This question was answered in the present experiment by comparing the running speeds to a nonrewarded LFA for *S*s having 12, 1, or 0 pellets on the MFA (Groups 12-0, 1-0, and 0-0, respectively). To the extent that running speed to this nonrewarded LFA is inversely related to the magnitude of reward on the MFA, a SimCE would have been defined in the form of a depression effect.

c. The simultaneous elation effect. To determine whether a simultaneous elation effect also occurred in this situation, running speed to the MFA was compared for *S*s in Groups 12-0, 12-1, and 12-12. To the

extent that MFA speeds were inversely related to magnitude of reward on the LFA, a SimCE would be defined in the form of an *elation* effect.

d. The effect of distribution of trials on the SimCE and SucCE. Although a SimCE in the form of a depression effect had been readily obtained in Experiments I, II, and III of the present report as well as in other research (e.g., Spear & Hill, 1965), one experiment from our laboratory conspicuously failed to demonstrate it (Spear & Pavlik, 1966). Spear and Pavlik employed the same procedure as in our other experiments (including the identical apparatus and *E*s as in Experiment II of the Spear and Hill paper) with the exception of the distribution of trials: Spear and Pavlik gave only one trial per day. Not only did they fail to obtain a simultaneous depression effect during Preshift, they found running speed to the LFA reliably greater for *S*s in Group 12-1 compared to Group 1-1, and they also found running speed to the MFA to be greater for Group 12-1 than for a group given 12 pellets on either alternative (which defined a simultaneous elation effect). They presented evidence that this was not due to the fact that the experiments employing more than one trial a day had resulted in animals which were differentially satiated on food and that their results were not due to any chance error of having unusually fast *S*s in their Group 12-1. Apparently, the employment of only one trial per day caused the difference.

It became clear that an experiment was needed in which similar conditions were employed but in which intertrial interval was varied. If intertrial interval were the critical factor, this fact should show up when more than one trial is given per day (assuming the differential in interval is sufficient to produce differential behavior).

Thus, several conditions were included in the present experiment in which treatment differed in terms of the magnitude of reward which appeared in the alternatives of the T maze. Orthogonal to the magnitude of reward variable, intertrial interval was varied: in each condition, one subgroup of *S*s received a 15-sec. intertrial interval and the other subgroup received a 15-min. intertrial interval between each of their six trials given in a single day. To improve comparability, all conditions were run an equal number of times in each of several replications of the experiment. However, basic concern was with several relatively disjoint questions asked within this experiment rather than with the complete factorial design as it eventually developed.

Method

Subjects and apparatus. The *S*s were 96 experimentally naïve female albino rats of the Sprague-Dawley strain, approximately 60 days old at the start of training and weighing 180–200 gm. The T maze, housing of *S*s, and running procedure (moving of cage rack outside the testing cubicle) were the same as in Experiment I. The inserts were also as used in Experiment I, with White always on the left and Brown always on the right.

Procedure. The prehandling and deprivation schedule was essentially the same as in Experiment I. The *S*s received ad lib access to water and were given a total of 10.4 gm. of food daily consisting of pellets and chow.

Experiment IV was run in four replications of 24 *S*s each. Each replication contained two *S*s from each of 12 groups, one of which was assigned Brown (right) as the MFA; the other, White (left) as the MFA. The 12 groups comprised a 6 × 2 factorial design varying Preshift magnitude of reward and intertrial interval. Half of the *S*s received their six daily trials in relatively rapid succession, being returned to their home cages for about 15 sec. between trials; the remainder were run in rotation with a resulting 12–15-min. intertrial interval. An individual *S* experienced the same intertrial interval throughout the entire experiment. The 12 groups were: 0-0-M, 0-0-S, 1-0-M, 1-0-S, 1-1-M, 1-1-S, 12-0-M, 12-0-S, 12-1-M, 12-1-S, 12-12-M, and 12-12-S. The first number indicates Preshift magnitude of reward (number of pellets) on the MFA, the second number indicates Preshift magnitude of reward on the LFA, M is massed trials (15-sec. intertrial interval), and S is spaced trials (12–15-min intertrial interval). During Postshift, all *S*s received one pellet on either alternative.

The running procedure was the same as in Experiment I, except that *S*s now received only one free trial per day. For *S*s in replications one and three, the first trial of each day was free, Trial 2 was forced to the side opposite that chosen on Trial 1, and Trials 3–6 were forced on a random basis, half to each alternative for each *S* each day. For *S*s on replications two and four, Trial 5 of each day was free, Trial 6 forced to the side opposite that chosen on Trial 5, and Trials 1–4 were

forced on a random basis, half to each alternative. Preshift, as in Experiment I, comprised Trials 1–48, (Days 1–8). Postshift was given through Trials 49–144 (Days 9–24).

Response measures. Only the committed speed and choices are reported for this experiment. Turning speed replicated previous results: less sensitivity than committed speed to the successive shift in reward, and a high correlation with choice. Turning speed may be a somewhat misleading measure anyway, particularly in a position-discrimination task. First, it exaggerates the difference between MFA and LFA speed because the initial turning motion of *S* toward the MFA may take place before the forcing door is seen. Moreover, the SimCE obtained with this measure may be inflated by the initial turn toward the MFA; by comparison with committed speed, it is clear that response strength to the MFA is exaggerated in this way when the turning-speed index is used. Although in terms of competing responses turning speed is as legitimate an index as any, even though this artifact does occur, we prefer to view the SimCE as primarily due to slower approach to the LFA rather than a consequence of greater incentive from such a specific competing alternative. Committed speed provides a somewhat purer measure in this respect. Finally, turning speed may provide a spuriously slow index of response strength in groups with equal magnitudes of alternative reward because this measure includes the time taken to VTE and otherwise resolve conflict at the choice point.

Results

Preshift choice behavior. As expected, the average preference for their MFA by those *S*s choosing between equal reward magnitudes (Groups 0-0, 1-1, and 12-12) was about equal throughout to that for their LFA. Therefore, analyses of choice behavior presented below concern only those *S*s receiving differential rewards in the alternatives (Groups 1-0, 12-0, and 12-1). A chi-square test on total number of choices of the MFA during Preshift revealed that reward magnitude reliably affected choice behavior. Groups 12-0 and 12-1, which did not differ, had more choices of the MFA during Preshift than did Group 1-0, $\chi^2(2) = 13.71$, $p < .001$. The overall mean probability of choosing the MFA was .87 for Group 12-0, .86 for Group 12-1, and .74 for Group 1-0. No reliable effect of trial spacing was obtained; the overall mean probability of choosing the MFA was .83 for *S*s given massed trials and .81 for *S*s given spaced trials. There was no interaction between trial spacing and reward magnitude.

It is perhaps worth mentioning that the relative and absolute preference for the MFA among the various experimental conditions was the same whether choice behavior was measured on the first trial or the fifth trial of each day. This has been the typical finding in experiments employing the present paradigm and is relevant to questions concerning both the effect of distribution of trials and the potential food satiation when more than one trial is given per day. The first trial of a day follows by more than 22 hours the immediately preceding trial, while the intertrial interval preceding the fifth daily trial is necessarily much less than 22 hours.

Change in behavior during early Preshift trials. The early growth of the discrimination may be assessed in terms of the progressive change in the differential of behavior directed toward the MFA compared with that directed toward the LFA. The specific analysis chosen in this case (a three-way mixed analysis of variance) compared the mean difference in MFA and LFA committed speeds within Groups 12-0, 12-1, and 1-0 on Preshift Days 1 and 2 with that on Preshift Days 3 and 4.

It was found that the extent to which the MFA speed was greater than the LFA speed was directly related to the difference in the alternative reward magnitudes. That is, from the greatest difference between MFA and LFA speeds to the least difference, the groups were ordered 12-0, 12-1, 1-0, $F(2,42) = 6.66$, $p < .005$. There was no reliable main effect of distribution of trials, nor did this variable enter into any reliable interaction. The difference between LFA and MFA speed was, of course, greater on Days 3 and 4 than on Days 1 and 2, $F(1,42) = 10.03$, $p < .005$, and there was a reliable interaction of reward magnitude with days, $F(2,42) = 10.12$, $p < .001$. This latter result reflected the increasingly more rapid discrimination formed in those groups with the greater differential in their alternative magnitudes of reward.

The same differences were obtained in

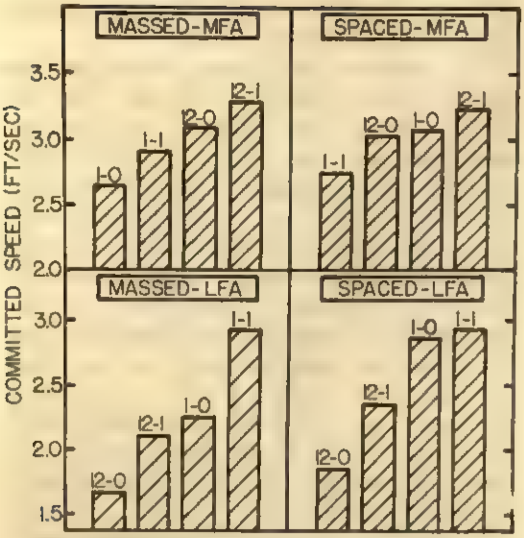

FIG. 11. These comparisons in terms of LFA speed represent a test for the simultaneous depression effect. Mean committed speed to the LFA and MFA is shown during the last two Preshift days (for massed and spaced Groups 12-1, 12-0, 1-1, and 1-0) in Experiment IV.

terms of mean committed speed on the final two Preshift days (see Figure 11). The difference between MFA and LFA speeds remained greater the greater the differential in alternative rewards, $F(2,42) = 5.42$, $p < .01$. Also at this point there was no effect of distribution of trials, and trial distribution did not interact with reward magnitude ($F < 1$ in each case).

Simultaneous elation effects at the end of the Preshift stage. Considering first the mean committed speed to the MFA, an analysis of variance including Groups 12-1, 12-0, 1-1, and 1-0 was performed with MFA reward magnitude, LFA reward magnitude, and distribution of trials as fixed orthogonal variables. The relevant comparisons may be seen in Figure 11 (MFA speed).

An elation effect would have been defined if MFA speed were greater the less the reward magnitude on the LFA (Groups 12-0 and 1-0 compared to 12-1 and 1-1). It was found that neither reward magnitude on the LFA, nor distribution of trials, nor any interactions produced reliable variance (F values ranged from .42 to 1.22 for all nonsignificant effects); thus, no elation effect was found in terms of this analysis. The only reliable source of variance in this case was greater speed to the MFA the greater the reward magnitude on the MFA, $F(1,56) = 5.27$, $p < .05$.

Committed speeds on the MFA were also examined by another analysis for evidence of an elation effect. Recall that a simultaneous elation effect would be defined if MFA speeds to a common MFA reward magnitude were inversely related to the magnitude of reward on the LFA. Thus, the MFA speeds for *S*s in Groups 12-0, 12-1, and 12-12 (see Figure 12) were compared under the two levels of trial distribution by an analysis of variance. Neither the effects of reward magnitude nor trial spacing approached significance at the .05 level—no elation effect occurred. The F value for the nonsignificant interaction, $F(1,42) = 2.61$, $p > .05$, was clearly inflated by the food-satiation effect which occurred in Group 12-12 when massed trials were given. Thus, it is concluded

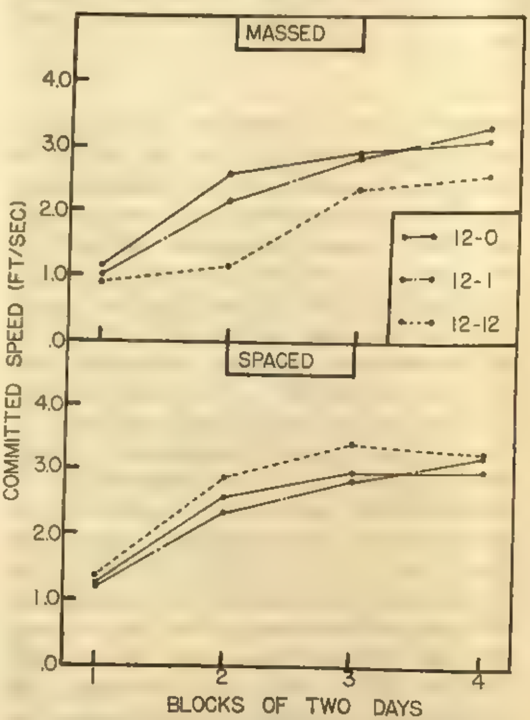

FIG. 12. This comparison represents a test for the simultaneous elation effect. Preshift committed speed to the MFA is shown for massed and spaced Groups 12-0, 12-1, and 12-12 in Experiment IV.

that no evidence for a simultaneous elation effect was obtained under these circumstances.

Simultaneous depression effects at the end of the Preshift stage. A basic aim of this experiment was to determine whether SimCEs occur in the conventional discrimination task involving a choice between some reward and no reward. With the present design, it was possible to go one step further and to examine the extent to which such a SimCE might be comparable in magnitude to the SimCE obtained when the choice is between a large and a small reward. The appropriate analysis included a comparison of LFA committed speeds by Groups 12-1, 12-0, 1-1, and 1-0 on the last 2 days of the Preshift stage (see Figure 11). Thus, the three orthogonal variables in this analysis were reward magnitude on the MFA, reward magnitude on the LFA, and distribution of trials.

Overall, the simultaneous depression effect did occur as speed to the LFA was less the greater the reward magnitude on the MFA, $F(1,56) = 15.51$, $p < .001$. Perhaps more important is the fact that this simultaneous depression effect did not differ whether the choice was between something and nothing or whether the choice was between something large and something small, as reflected by the absence of interaction between MFA and LFA reward magnitude, $F(1,56) = .06$. Speed to the LFA was greater when one pellet was the reward in the LFA than when no pellets were present, but neither the effect of distribution of trials nor any remaining interaction approached statistical reliability (Fs ranged from .06 to 1.88). These results imply that the SimCE is present when *S* chooses between some nominal reward and no reward. Furthermore, it appears that the magnitude of this SimCE does not deviate from that obtained in a choice between a large and a small reward. Finally, the absence of interaction between MFA reward and distribution of trials shows that our prediction in this respect was incorrect: the SimCE does not decrease with longer intertrial interval.

Another way of investigating the SimCE when the discrimination is between stimuli associated with some reward versus no reward is by analyzing the committed speeds on the last 2 days of the Preshift stage of Groups 12-0, 1-0, and 0-0 orthogonal to the distribution-of-trials variable. This analysis revealed a slight complication. The reliable effect of reward magnitude, $F(2,42) = 3.89$, $p < .05$, largely reflected the uniformly greater LFA speed by *S*s in Group 1-0 compared with *S*s in Group 12-0. However, the mean LFA speeds in Group 0-0 were only slightly greater than those in Group 12-0, and this was the case only under conditions of spaced trials. This effect was not great enough to result in a reliable interaction (Distribution of Trials × Reward Magnitude) ($F < 1$), though it did contribute to the significantly greater LFA speeds overall for *S*s run under spaced trial conditions, $F(1,42) = 5.08$, $p < .05$. It is not unlikely that a condition such as that of Group 0-0, in which no portion of behavior is under the control of food reward, may not provide an adequate baseline for the estimation of CEs in *S*s whose behavior is under the control of food reward to at least some extent.

Choice behavior during the Postshift stage. Since all *S*s in Groups 12-0, 12-1, and 1-0 chose the MFA on the last free trial of the Preshift stage and also chose this alternative during each free trial on the first 2 days of Postshift, the change in choice behavior was analyzed in terms of the trial on which the formerly LFA was first chosen following the shift in reward magnitude in the MFA. Preshift reward magnitude made little difference in this respect. The mean Postshift trials of the first LFA choice were 8.19, 9.19, and 8.25, respectively, for Groups 12-0, 12-1, and 1-0; and the respective mean probabilities of choosing the former MFA throughout Preshift were .79, .77, and .76.

However, those *S*s given spaced trials persevered in choosing the formerly MFA to a considerably greater extent than those given massed trials (see Figure 13). The reliability of this was substantiated by a Mann-Whitney U test in terms of the trial on which the first LFA choice was made during Postshift, $U = 30.25$, $p < .001$.

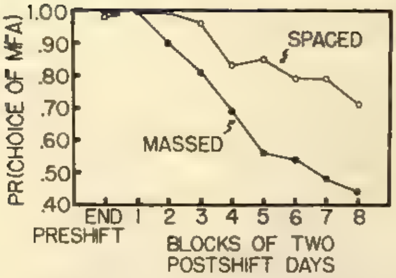

FIG. 13. The course of choice behavior following the shift in reward as a function of distribution of trials. Scores for Groups 12-0, 12-1, and 1-0 are combined.

Combining the differentially rewarded groups, it was found that the mean free Postshift trial on which the first LFA choice was made was 5.92 for *Ss* given massed trials and 11.08 for *Ss* given spaced trials. There was no interaction between Preshift reward and distribution of trials.

It is notable that the basic relationship between distribution of trials and persistance in choosing the former MFA after the reward shift did not differ whether only the first trial (which was preceded by a 24-hr. intertrial interval in both the massed and spaced conditions) or the fifth trial of each day was considered. This suggests that the greater perseverance associated with more widely spaced trials cannot be dismissed as a simple consequence of a greater tendency under massed trials to alternate stimuli between the fourth and fifth daily Postshift trials. Also, this fact is not likely a consequence of the greater number of errors by the spaced-trial groups during Preshift (cf. D'Amato & Jagoda, 1960) since number of "errors" was, in fact, equated for all *Ss* by employing forced trials.

Committed speed to the MFA following the shift in reward magnitude: Test for SucCE. There are several ways to evaluate the SucCE and several questions concerning its occurrence. First consider committed speed to the MFA during the first 4 days of Postshift. With this measure, an analysis was completed employing the basic factorial design, which included Groups 1-0, 1-1, 12-0, and 12-1. These speeds are shown in Figure 14, combining massed- and spaced-trial groups. Recall that three orthogonal variables are MFA reward magnitude, LFA reward magnitude, and distribution of trials. The only reliable source of variance was found to be contributed by the SucCEs: committed speed to the MFA was slower for those *Ss* previously having the larger MFA reward magnitude, $F(1,56) = 10.41$, $p < .001$. There was some tendency toward greater MFA speed for those *Ss* (Groups 12-1 and 1-1) which previously had the larger LFA reward, but this effect did not attain statistical reliability, $F(1,56) = 2.94$, $p > .05$. None of the other five sources of variance approached statistical significance (F values ranged from .01 to 1.79). Thus the SucCE (depression) was defined and provided the only reliable source of variance in this analysis.

Test for SucCE (depression) on the MFA as a function of reinforcement history on the LFA. The SucCE may be evaluated by analyzing the effects of Preshift reward magnitude and distribution of

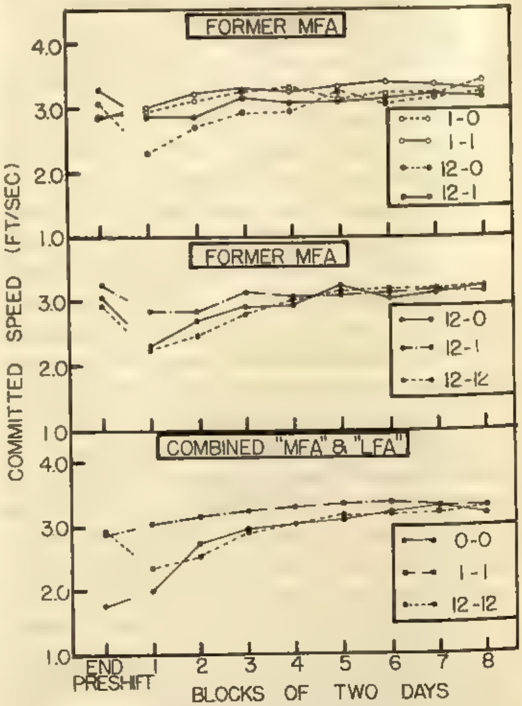

FIG. 14. These three comparisons represent tests of the SucCE. Postshift committed speeds are shown to the formerly MFA, summed over massed- and spaced-trial conditions in Experiment IV.

trials as a function of Postshift experience. When committed speeds to the alternative shifted from 12 pellets to 1 pellet (Groups 12-0, 12-12, 12-1) are thus compared to corresponding speeds for the baseline control Group 1-1, slower overall speed by a group which formerly received 12 pellets would define a SucCE. A SucCE would also be present if the interaction between Preshift reward magnitude and Postshift days contributed significant variance, assuming the interaction were such that the group which previously had the larger reward increased their initially depressed speed relative to the lesser change by Group 1-1. The general Postshift performance of these four groups may be seen in Figure 15.

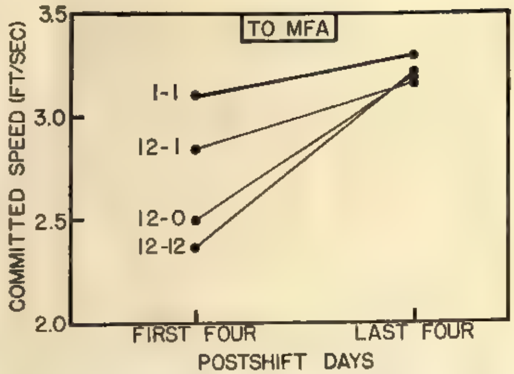

FIG. 15. Committed speed to the MFA during initial and terminal portions of Postshift: magnitude of the SucCE on the MFA as a function of prior reward on the LFA.

The MFA committed speeds of Groups 12-0 and 1-1 were compared over eight blocks of 2 Postshift days. A SucCE was found summing over blocks of Postshift days—speed to the formerly MFA was less for Group 12-0, $F(1,28) = 6.26$, $p < .025$. This SucCE was also reflected by the significant interaction of reward magnitude by blocks, $F(7,196) = 5.84$, $p < .001$. Figures 14 and 15 show that this interaction was a consequence of the increase in speed over blocks by Group 12-0 following the earlier SucCE. The only remaining significant effect from this analysis (the triple interaction, $F(7,196) = 3.80$, $p < .001$) is attributed entirely to the inflated mean speed in Group 1-1 with spaced trials.[2]

Independent analyses of variance comparing the MFA speeds of Groups 12-0 and 1-1 on each block of 2 Postshift days determined that the effect of Preshift reward magnitude no longer contributed reliable variance after the seventh and eighth days of Postshift. Thus, the committed speed to the MFA of *S*s in Group 12-0 apparently adjusted to the level of their baseline control by about the ninth and tenth days of Postshift.

It appeared that the SucCE described above for *S*s in Group 12-0 was a more powerful effect than had been obtained previously under conditions comparable to those of Group 12-1. In fact, the present Group 12-1 did not show a statistically reliable SucCE either in terms of slower mean committed speed (MFA) compared to Group 1-1, $F(1,28) = 2.57$, $p > .10$, or in terms of the significant interaction between the effects of reward magnitude and blocks of Postshift days, $F(7,196) = .56$. However, it should be noted that the tendency for a SucCE in Group 12-1 was present numerically though not statistically reliable at the .05 level. Since this tendency has been found under these conditions in a total now of six experiments—sometimes attaining statistical reliability (at the .05 level) and sometimes not—the

[2] It is necessary to comment on the running speed of Group 1-1 which was given spaced trials. The mean speed by this group was consistently greater than that by any other group throughout the Postshift stage. It is our belief that this is a spurious result—a matter of sampling error. There are several bases for this contention. First, in an absolute sense, the Postshift speed by this group was greater than that previously obtained under these conditions with the same apparatus and under conditions of nearly comparable intertrial interval. Second, as can be seen to some extent in the LFA speeds shown in Figure 16, this group showed considerable increase from the end of Preshift throughout most of Postshift; this has not been found previously under these conditions. Finally, in several previous experiments of this kind, we have always found that the mean speed to the MFA by groups shifted in reward magnitude adjusts to a common level—that of the baseline control—by the end of Postshift. This was not the case in this experiment, but only because of the faster running by Group 1-1 given spaced trials. Thus, we conclude that the Postshift speed of this group is inflated and overestimated due to chance factors.

danger of a Type I error is probably not too great to permit the conclusion that a SucCE was present in Group 12-1. The test of this tendency for this effect to be weaker in Group 12-1 than in Group 12-0 is presented below.

The SucCE could also be defined in terms of the Postshift performance of *S*s in Group 12-12. This group had reliably slower speeds than were found in Group 1-1, $F(1,28) = 8.91$, $p < .01$; thus, the successive depression effect also occurred in Group 12-12. The fact that this effect occurred most strongly during the early stages of Postshift is reflected in the reliable interaction of Preshift reward magnitude and Postshift blocks, $F(4,112) = 2.54$, $p < .05$. Independent analyses of variance over each block of 2 Postshift days suggested that the speeds of Group 12-12 had adjusted to the level of Group 1-1 by the fifth and sixth days of Postshift, when the effect of Preshift reward magnitude was no longer reliable ($p > .05$). The effect of distribution of trials, though inflated by the chance occurrence of Group 1-1 faster speeds under spaced conditions, did not attain statistical reliability in this analysis, $F(1,28) = 3.00$, $p > .05$, and all other sources of variance yielded *F*s less than 1.

The results of the above analyses suggest that the successive depression effects obtained were stronger in Groups 12-0 and 12-12 than in Group 12-1. These groups differed only in terms of LFA reward experienced during Preshift. However, the prior reinforcement conditions were such that Group 12-12 never had experience with a SimCE nor the specific magnitude of reward (one pellet) presented during Postshift, while Group 12-0 had experienced the former but not the latter. In order to determine whether only one or both of these factors had caused the difference from Group 12-1 in terms of adjustment to the Postshift reward, committed speed to the MFA was compared for these groups during the Postshift stage (employing the Reward Magnitude × Distribution of Trials × Blocks of 2 Postshift Days analysis of variance as above).

The results supported the conclusion that both Groups 12-12 and 12-0 were more affected by the shift from 12 pellets to 1 pellet on the MFA than was Group 12-1 and in the direction of a greater SucCE. The Postshift committed speeds of Groups 12-12 and 12-0, however, did not differ. The results of the analysis of variance comparing Group 12-1 with 12-12 and that comparing Group 12-1 with 12-0 were essentially the same. In both cases the only reliable source of variance, with the exception of the obvious effect of Postshift days, was the interaction between reward magnitude and blocks of Postshift days—for 12-1 versus 12-12, $F(7,196) = 2.81$, $p < .01$; for 12-1 versus 12-0, $F(7,196) = 3.18$, $p < .005$. This interaction reflected the fact of eventual adjustment to the same level of MFA committed speeds (at about Postshift Days 6–10) for all groups, but with both of Groups 12-12 and 12-0 adjusting from a lower level early in Postshift than was the case in Group 12-1. No other effects from these analyses approached statistical reliability—the *F* values ranged from .01 to 1.38. When the same analysis compared Groups 12-0 and 12-12, neither the effect of reward magnitude nor the interaction between reward magnitude and blocks of Postshift trials attained statistical reliability ($F < 1$ in each case). It is concluded that the shift in MFA reward did not differentially affect *S*s in Groups 12-0 and 12-12, but in both cases the effect was greater than that found in Group 12-1.

Test for SucCE (elation) in group shifted from zero nominal reward to small reward. To the extent that the mean speed in Group 0-0 overshot that in Group 1-1, a successive elation effect would have been defined. However, the test—a three-way analysis of variance comparing committed speeds of Group 1-1 with those of Group 0-0 under conditions of massed or spaced trials and over the first five blocks of 2 Postshift days—revealed that no successive elation effect had occurred. In fact, *S*s having the 1-1 reward condition maintained faster running speed than those in the 0-0 condition—when scores were combined across blocks of Postshift days, $F(1,28) = 17.42$, $p < .001$. Thus, it is con-

cluded that no successive elation effect occurred.

Running speed to the LFA during the Postshift stage. Subsequent to a simultaneous depression effect, a reduction in the magnitude of the contrastingly large MFA reward results in the adjustment of the previously depressed performance on the LFA: these LFA speeds increase toward the level of expected performance, as defined by the baseline control group (see, for example, Experiment I and II of the present report; Spear and Hill, 1965). This adjustment was investigated here as a function of reward magnitude on the MFA during Preshift, reward magnitude on the LFA during Preshift, and distribution of trials. Of course, only the preformance of the differentially rewarded groups is relevant here, in relation to the baseline control Group 1-1 and in relation to each other. These speeds (see Figure 16) were analyzed in several ways, but only a few results warrant consideration here.

First, the direction of two interactions—between MFA reward magnitude and blocks of Postshift days, $F(1,56) = 11.32$, $p < .001$, and between LFA reward magnitude and blocks of Postshift days, $F(1,56) = 6.30$, $p < .025$—revealed two clear and related facts. It showed that the increase in LFA committed speeds during Postshift was greater the larger the Preshift reward on the MFA and the smaller the Preshift reward on the LFA. Thus, these relationships reflect the combined effects of absolute *and* relative reward magnitude experienced on the LFA prior to the shift. Second, *S*s which previously had no reward on the LFA showed a greater increase in LFA speeds the larger the previous MFA reward (for the interaction between the MFA reward, LFA reward, and blocks of Postshift trials, $F(1,56) = 6.34$, $p < .025$). This appeared to be a consequence of the initially slower LFA speeds by Group 12-0 relative to Group 12-1 in comparison to the smaller difference between Groups 1-0 and 1-1. It is clear that this effect of MFA reward is one reflection of the SimCE.

Finally, since a powerful simultaneous depression effect had been found in Group 12-0 relative to Group 1-0 during Preshift, we may ask how this effect was altered during the initial stages of Postshift after reward had been decreased in the MFA *and* increased in the LFA for Group 12-0, but only increased in the LFA for Group 1-0. A three-way analysis of variance was completed comparing the effects of Preshift reward magnitude and distribution of trials on the last blocks of 2 Preshift days and the first block of 2 Postshift days. Although the overall simultaneous depression effect was, of course, reliable, $F(1,28) = 19.28$, $p < .001$, no other effects approached statistical significance. This suggests that the effects of the reward shift on the first 2 days of Postshift were not great enough in terms of LFA committed speed for these groups to alter the effects that had obtained during the terminal stages of Preshift.

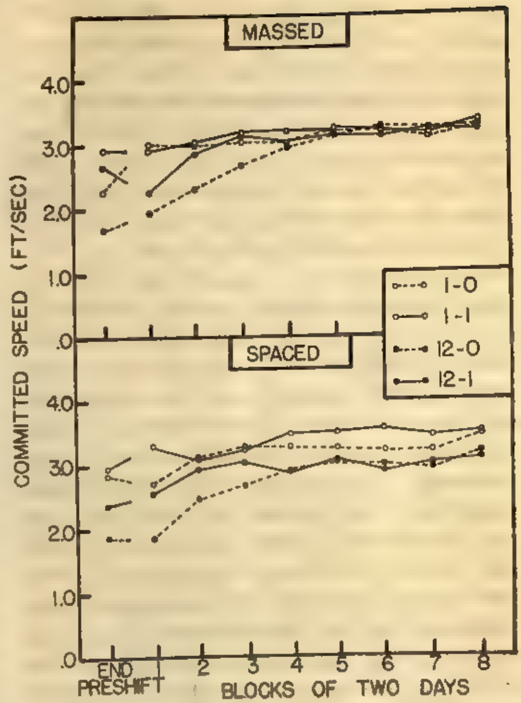

FIG. 16. Postshift committed speed to the LFA for massed and spaced Groups 12-1, 12-0, 1-1, and 1-0 in Experiment IV.

Discussion

The most important results in this experiment concern differences in committed

speed; there was again no evidence for a SucCE in choices. This was true even for Group 12-0, which was shifted both to larger LFA and smaller MFA rewards. As usual, greater preference for the MFA was obtained the greater the differential in the alternative reward magnitudes, and this preference was never reversed when the alternative rewards became equal.

Perhaps the one interesting result in choice behavior was the greater persistence in choice of the formerly LFA following the reward shift for those *S*s run under conditions of spaced trials. A similar result has been reported by Clayton (1964) for rats in a discrimination-reversal task (although the analogy with the present paradigm is not perfect, the same processes may be involved). Clayton found that *S*s run under a 20-min. intertrial interval were more reluctant to discontinue their choice of the original MFA than were *S*s run under an intertrial interval of 10 sec. or 3 min. He interpreted this in terms of greater forgetting during the longer intertrial interval. Both his data and interpretation are compatible with similar runway phenomenon (Hill, Erlebacher, & Spear, 1965; Spear, 1965; Spear, Hill, & O'Sullivan, 1965) and also would fit the present results rather well. It is surprising that committed speed was not similarly affected by the differential intertrial interval. There is, however, evidence that running speed is less susceptible to the effects of a retention interval than is behavior (such as choice) that reflects the efficiency of the discrimination (Spear, Hill, & Cotton, 1962).

Simultaneous contrast effects. Although the simultaneous depression effect occurred as usual, no evidence could be found for a simultaneous elation effect. This is not inconsistent with the results obtained in the straight runway with the SucCE paradigm. An elation effect, defined in terms of reliably greater speed in a group shifted from small to large reward compared with a baseline group which has always had the large reward, has rarely, if ever, appeared in the literature. For example, neither the classic experiments of Crespi (1942) nor Zeaman (1949) obtained results which satisfied this definition of an elation effect. As Crespi himself pointed out, his *S*s which showed the "elation effect" had *originally* received large reward prior to being rewarded with small, then shifted *again* to large, reward; and Zeaman lacked definitive control groups. Perhaps the absence of an elation effect is a consequence of physiological limits imposed upon running speed in rats. If not, this fact remains a distinct problem for perceptual interpretations of CEs (e.g., Bevan, 1963).

Another finding established the occurrence of a SimCE (a "depression" effect) in the standard discrimination paradigm in which *S* chooses between some and no reward. This conclusion was based on the finding of slower running speed to zero reward for *S*s in Group 12-0 compared with those in Group 1-0; it was not so unequivocal when the behavior of Group 0-0 was considered as a baseline control. However, there are several reasons why the behavior of *S*s in Group 0-0 (who never experienced nominal reward during the Preshift stage) may not be considered an appropriate baseline. First, no portion of the behavior of these *S*s was under the control of nominal reward during Preshift; this was not the case for *S*s in Groups 12-0 and 1-0. Second (and this may be a reflection of the first point), there was a distinct tendency for faster running under spaced-trial conditions by *S*s in Group 0-0, and this was the case in only one group (12-12) which received reward during Preshift. Furthermore, this behavior of *S*s in Group 12-12 could be attributed to a factor obviously not present in Group 0-0—greater food satiation under conditions of massed compared with spaced trials. Finally, there were observations by *E* that the behavior of *S*s which had never received food reward in the experimental situation was qualitatively distinct. This behavior may be described as hyperactive, hyperreactive, skittish, and highly variable. In the course of running some 8 or 10 experiments in our laboratory which have included similar conditions, we have invariably obtained the same (unsolicited) report from many different *E*s.

The third important result was the fact that distribution of trials per se had no

reliable effect on the SimCE. Of course, this conclusion is limited to the range of intertrial interval employed here. However, this range often has been shown to produce substantial differences in other behaviors (e.g., Clayton, 1964; Spear, 1965).

The lack of an effect of intertrial interval is important because of the results obtained by Spear and Pavlik (1966) with this paradigm under conditions of one trial per day. Spear and Pavlik did not obtain a simultaneous *depression* effect. The critical feature appeared to be the fact that only one trial per day was given and the intertrial interval (24 hr.) was, therefore, much longer than that employed when the SimCE (depression) was obtained. The present experiment has shown, however, that similar differential effects probably cannot be obtained simply by varying the intertrial interval within a daily session of trials. It is still possible, of course, that a greater range of intertrial interval (for example, 15 sec. versus several hours) may produce the implied interaction. Indeed, in terms of numerical differences, the SimCE (12-1 versus 1-1) was slightly greater with 15-sec. than 15-min. intertrial interval (see Figure 12); and 15 min. is considerably less than 24 hr. However, the complete absence of statistical reliability in this case—and the opposite numerical result when Groups 12-0 and 1-0 defined the SimCE—make this possibility seem less likely.

Another possibility is that the specific processes responsible for the effects obtained in the Spear-Pavlik experiment were somehow correlated with some aspect of the operation of presenting one trial per day. However, the specific nature of these processes if they exist, is not at all clear.

Successive contrast effects. Following the shift from 12 pellets to 1 pellet in the MFA, the effect on behavior early in Postshift was determined by the nature of the alternative rewards experienced prior to the shift. Those *S*s which had not experienced the Postshift magnitude of reward before the shift slowed their running speed to a greater extent after the shift. Specifically, the results indicated that the SucCE for *S*s in Group 12-0 was about equal to that in Group 12-12, but greater in both of these groups than in Group 12-1. Now it is not difficult to understand that the effect of the shift should be greater in Group 12-12 than in Group 12-1. In fact, both a perceptual and an emotional interpretation of CEs apparently can accommodate this fact with ease, in contrast to the SucCE in Group 12-0.

In terms of the adaptation level theory proposed by Bevan (1963), it is quite clear that the indifference point (defined as that reward magnitude judged neither large nor small but medium or neutral) would be lower for *S*s pooling the occurrence of 12 pellets and 1 pellet over trials (Group 12-1) than for *S*s receiving no reward magnitude in the experimental situation other than 12 pellets (Group 12-12). From this viewpoint, the Postshift reward magnitude of one pellet, therefore, would appear smaller to *S*s in Group 12-12 than to *S*s in Group 12-1 when judged against their respective indifference points. This perceptual interpretation would have predicted the greater effect of the reward shift in Group 12-12 in view of Bevan's (1963) assertion that, "Overall, however, it would appear that speed of running, like maze performance, varies with the apparent, in contrast to the physical, strength of the reinforcing agent. [p. 27]."

Although the interpretation of CEs by emotional terms has been perhaps less explicit, Bower (1961) has provided a clever application of frustration theory. His interpretation viewed the SimCE as resulting "... from a conflict between anticipation of reward (r_g) and the anticipation of frustration in the S— (LFA) goal box [p. 199]." The frustration elicited in the LFA goal box was assumed to occur, at least in the intermediate stages of discrimination training, "... when the larger amplitude r_g established in S+ occurs in S— through stimulus generalization [p. 199]." It is not inconceivable that this interpretation might also be applied to the SimCE found after discrimination had been perfect for some time (as measured by the independent choice measure). The resolution of this conflict being rewarded in

Preshift (*Ss* in Group 12-1 nearly always completed their run to the one pellet available in the LFA goal box) could account for the lesser effect of the shift to one pellet in the MFA for this group. By analogy with the interpretation of the partial reinforcement effect in extinction (Amsel, 1962) the *Ss* in Group 12-1 had their $r_f - s_f$ conditioned to running via their eventual approach to the (frustrating) one pellet during the Preshift stage and thus would be expected to respond in the same way to s_f, that is, to run correspondingly, when r_f was elicited in the MFA during Postshift. Since the Preshift conditioning of $r_f - s_f$ obviously was not present in Group 12-12, this group would be expected to be more disrupted by the "frustrating" occurrence of one pellet after the shift.

Apparently, though, both theories have difficulty explaining Postshift behavior on the MFA by *Ss* in Group 12-0. In every other respect, their behavior was very much like that of Group 12-1. This included the occurrence of the SimCE in Group 12-0 relative to Group 1-0, a fact which implies that when these *Ss* were rewarded on the MFA they responded to the LFA as if the zero reward were a point on the continuum of reward magnitude. That is to say that *Ss* in Group 12-0 did not respond to the zero reward with the same absolute running speed as *Ss* in Group 1-0. Rather, they responded to zero reward as if it were a smaller reward than that obtained on the MFA. Also as in Group 12-1, the Group 12-0 *Ss* nearly always eventually completed their run to the nonrewarded LFA goal box, even after the discrimination between the MFA and LFA had been perfect for some time. Thus, from either Bevan's concept of reinforcement pooling or an interpretation in terms of conditioned $r_f - s_f$, it would appear that *Ss* in Group 12-0 would be expected to perform more like *Ss* in Group 12-1 than like *Ss* in Group 12-12.

Obviously it cannot be certain that the conditions of Group 12-0 produce the same Postshift behavior in the MFA as would be found under conditions like Group 12-12; probably it does not. The difficult fact for theory, however, is that the behavior of *Ss* in Group 12-0 differed in the same direction and to about the same extent as the behavior of *Ss* in Group 12-12. Bevan (1963) could account for this only by asserting that zero reward is qualitatively different from a very small reward. This may be true in certain instances (for example, when none of *S*'s behaviors are under the control of nominal reward), but this does not appear to apply to the LFA behavior of *Ss* in Group 12-0; at least not in view of the SimCE which they exhibited. Frustration theory would seem to have the greater potential in this respect by appealing to the specificity of r_g. Perhaps the critical feature in the 12-0 group is the absence of conditioned r_g for one pellet prior to the reward shift. The *Ss* in Group 12-1 already had such an r_g conditioned to certain stimuli in the maze situation and could thus make use of it, at least in generalized form, in the MFA. But *Ss* in Group 12-0 were required to "start from scratch" with the conditioning of this new r_g. It may be assumed that $r_f \rightarrow s_f$ was experienced equally in the MFA by these groups. Thus, the slower initial Postshift speeds and slower adjustment to the expected level of performance would be accounted for in Group 12-0 relative to Group 12-1.

Perhaps the most likely possibility is that the greater SucCE found in Groups 12-0 and 12-12 compared with 12-1 is not the result of "contrast effect" at all but of generalization decrement. A most reliable occurrence is reduced running speeds by rats contingent upon some change in the stimulus situation. The change in reward magnitude, especially to a magnitude not previously experienced, would constitute such a change. Although the SimCE could not be accounted for by generalization decrement, it is likely that some, if not all, of the SucCEs which have been measured may be due to this factor. Spear and Spitzner (1965) have cited several diverse sources of evidence which point to generalization decrement as a primary contributor to the (operationally defined) successive depression effect. Surely, this factor cannot be discounted as a major source of

the variance in the *measured* SucCE so long as a successive elation effect is not reliably demonstrated under comparable circumstances.

GENERAL DISCUSSION

This series of experiments has provided some answers relevant to the three general points mentioned in the introduction to this paper. Consideration of these points is given below.

SimCE and SucCE Compared

First, a rough comparison of the magnitude of the SimCE and SucCE was possible from Experiment IV. In Group 12-12 these effects were not contaminated by prior experience with the Postshift reward and/or the SimCE. Group 12-0 had prior experience with the SimCE but not with the Postshift reward. In general, it did not appear that the SucCE found in Groups 12-12 and 12-0 was appreciably different than the SimCE found in Groups 12-1 and 12-0. This is a surprising fact in view of the predictably greater SimCE due to more available comparisons of the contrasting rewards, lesser retention interval from one reward to the other, etc. For example, the results in terms of the Placed groups of Experiment I had suggested that the SimCE was the more robust phenomenon. It was noted in Experiment IV, however, that the more typical quantitative similarity could be misleading since the SucCE may be due to factors other than those which contribute to SimCEs. For example, stimulus generalization decrement may contribute heavily to the former but probably not to the latter.

Is the SucCE Ubiquitous in Its Effect on S's Responses?

A second point raised by the introduction was the question of the extent to which a CE associated with a given stimulus affects *S*'s responses to other stimuli. That is, to what extent does the CE share the ubiquitous character found in the partial reinforcement effect on extinction?

In terms of committed speed immediately following the shift, performance on the LFA was completely influenced by the reward decrement on the MFA—as committed speed decreased on the MFA, an immediate and identical decrease occurred on the LFA. This fact is illustrated in Figure 17. It was established that the trial-by-trial decline in speed during the first Postshift day was greater for *Ss* in Group 12-1; for the Groups × Trials interaction, $F(2,112) = 9.43$, $p < .001$. This is to be expected since reward was decreased only in Group 12-1, not in Group 1-1. Of greater importance was the apparent fact that the decline in LFA speed did not differ from that on the MFA; this was supported by the lack of a reliable interaction between Groups, Alternatives (MFA versus LFA), and Trials, $F(2,112) = 1.87$, $p < .25$. Thus, it appeared that a performance decrement on the LFA occurred which was parallel to that on the MFA, even though reward was reduced only on the MFA. Nevertheless, it was possible that the uniform within-days decline in performance by Group 12-1 was not unique to the *reward reduction* on the MFA. To test this, the performance by *Ss* in Group 12-1 was compared within the last day of Preshift and during the first day of Postshift. The interaction between Days and Trials within Days, $F(2,56) = 2.83$, $.05 < p < .10$, reflected the lack of variation in perform-

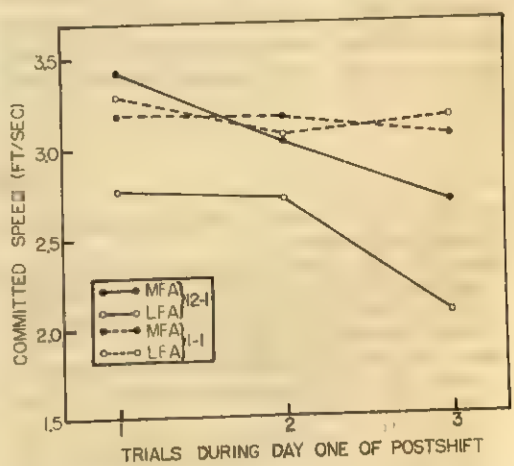

FIG. 17. Performance within the first day of Postshift. Data from Experiments I and II are combined; each of Groups 1-1 and 12-1 includes 32 *Ss*.

ance within the final Preshift day in contrast to the uniform decrement obtained *after* the MFA reward was decreased. Moreover, the three-way interaction among Trials, Alternatives, and Days did not approach statistical reliability; this supports the contention that the common within-day trend of MFA and LFA performance did not differ before and after the shift.

Therefore, it is concluded that following the reduction in MFA reward, the decrease in MFA committed speed is accompanied by an equivalent decrease in committed speed on the LFA—in spite of the fact that reward was not changed on the LFA.

A corresponding effect was not obtained in terms of turning speed; the eventual MFA decrease was accompanied by appropriate LFA readjustment (increase). During the first day of Postshift, turning speed remained quite constant from trial to trial on both the LFA and MFA. This fact argues against a conclusion that several or all of *S*'s responses, including some clearly distinct from that response directly instrumental to the altered reward, are affected by a single-reward change. Perhaps the common decline in LFA and MFA committed speed was due to the number of common stimulus elements in these "committed" portions of the maze (the comparison available at the choice point probably reduces the importance of this factor in terms of turning speed). It has been argued that this is not the case in the response-ubiquitous partial reinforcement effect (PRE) (Spear & Pavlik, (1966), but it must remain a possible explanation of the present results until further tests are made.

CE in Choice Behavior

Finally, a CE in terms of choice behavior never occurred. Three factors work against its occurrence. These factors include the spatial-temporal ordering of occurrence of the CE from the goal back toward the starting point (e.g., Spear & Spitzner, 1965; Vogel, Mikulka, & Spear, 1966) in combination with the transient nature of the SucCE (Gonzales, Gleitman, & Bitterman, 1962) and the change in LFA performance paralleling that on the MFA, when reward is shifted only on the MFA (see above). It is possible that, in any choice situation, the influence of the CE may be so weakened by the time the effect works its way back to the choice point that the influence on *S*'s behavior there is relatively minor. Moreover, assuming preference for the former LFA is contingent upon greater response strength to that side, it would appear that following the reward shift, either the SimCE must recover before the SucCE or the SucCE must have the greater effect, if a CE in choices is to be obtained. Neither of these events occurred in the present paradigm. Perhaps the extent of the effect of a reward shift on choice is eventual adjustment to the expected (baseline) level of preference.

REFERENCES

BARNETT, S. A. The rat: A study in behavior. Chicago: Aldine, 1963.

BEVAN, W. The pooling mechanism and the phenomenon of reinforcement. In O. J. Harvey (Ed.): *Motivation and social interaction.* New York: Ronald Press, 1963. Pp. 18–34.

BIRCH, D. Extended training and incentive shifts. Paper read at Psychonomic Society, Niagara Falls, October, 1964.

BOWER, G. H. A contrast effect in differential conditioning. *Journal of Experimental Psychology*, 1961, **62,** 196–199.

BOWER, G. H., & TRAPOLD, M. A. Reward magnitude and learning in a single-presentation discrimination. *Journal of Comparative and Physiological Psychology*, 1959, **52,** 727–729.

CAPALDI, E. J., HART, D., & STANLEY, L. R. Effect of intertrial reinforcement of the aftereffect of nonreinforcement and resistance to extinction. *Journal of Experimental Psychology*, 1963, **65,** 70–74.

CRESPI, L. P. Quantitative variations of incentive and performance in the white rat. *American Journal of Psychology*, 1942, **55,** 467–517.

CLAYTON, K. N. T-maze acquisition and reversal as a function of intertrial-interval. Paper read at Southeastern Psychological Association, Gatlinburg, 1964.

COLLIER, G., & MARX, M. H. Changes in performance as a function of shifts in the magnitude of reinforcement. *Journal of Experimental Psychology*, 1959, **57**, 305–309.

D'AMATO, M. R. & JAGODA, H. Effects of extinction trials on discrimination reversal. *Journal of Experimental Psychology*, 1960, **59**, 254–160.

GONZALEZ, E. C., GLEITMAN, H., & BITTERMAN, M. E. Some observations on the depression effect. *Journal of Comparative and Physiological Psychology*, 1962, **55**, 578–581.

GOODRICH, K. P. Running speed as a function of sucrose concentration in a prior free-drinking period. *Psychological Reports*, 1962, **11**, 528–530.

GOODRICH, I. P., & ZARETSKY, H. Running speed as a function of concentration of sucrose incentive during pretraining. *Psychological Reports*, 1962, **11**, 463–468.

HILL, W. F., ERLEBACHER, A., & SPEAR, N. E. Reminiscence and forgetting in a runway. *Journal of Experimental Psychology*, 1965, **70**, 201–209.

KNARR, F. A., & COLLIER, G. Taste and consummatory activity in amount and gradient of reinforcement functions. *Journal of Experimental Psychology*, 1962, **63**, 579–588.

PEREBOOM, A. C. A note on the Crespi effect. *Psychological Review*, 1957, **64**, 263–264.

SPEAR, N. E. Choice between magnitude and percentage of reinforcement. *Journal of Experimental Psychology*, 1964, **68**, 44–52.

SPEAR, N. E. Proactive interference effects of initial nonrewarded trials. Paper read at Midwestern Psychological Association, Chicago, 1965. (a)

SPEAR, N. E. Replication report: Absence of a successive contrast effect on instrumental running behavior after a shift in sucrose concentration. *Psychological Reports*, 1965, **16**, 393–394. (b)

SPEAR, N. E., & HILL, W. F. Choice behavior and accompanying speed measures after a change in nominal reward: Simultaneous- and successive-contrast effects. Paper read at Psychonomic Society, Niagara Falls, 1964.

SPEAR, N. E., & HILL, W. F. Adjustment to new reward: Simultaneous- and successive-contrast effects. *Journal of Experimental Psychology*, 1965, **70**, 510–519.

SPEAR, N. E., HILL, W. F., & COTTON, J. W. T-maze performance as a function of percentage reinforcement and interval between acquisition and extinction. Paper read at Psychonomic Society, St. Louis, 1962.

SPEAR, N. E., HILL, W. F., & O'SULLIVAN, D. J. Acquisition and extinction after initial trials without reward. *Journal of Experimental Psychology*, 1965, **69**, 25–29.

SPEAR, N. E., & PAVLIK, W. B. Percentage of reinforcement and reward magnitude effects in a T-maze: Between and within subjects. *Journal of Experimental Psychology*, 1966, **71**, 521–528.

SPEAR, N. E.. & SPITZNER, J. H. Characteristics of simultaneous and successive contrast effects of reward magnitude. Paper read at the Psychonomic Society, Chicago, 1965.

SPENCE, K. W. *Behavior theory and conditioning.* New Haven: Yale University Press, 1965.

VOGEL, J. R., MIKULKA, P. J., & SPEAR, N. E. Effect of interpolated extinction and level of training on the "depression effect." *Journal of Experimental Psychology*, 1966.

ZEAMAN, D. Response latency as a function of the amount of reinforcement. *Journal of Experimental Psychology*, 1949, **39**, 466–483.

(Received December 31, 1965)

Vol. 80, No. 11 Whole No. 619, 1966

Psychological Monographs: General and Applied

FRUSTRATION AND SECONDARY REINFORCEMENT CONCEPTS AS APPLIED TO HUMAN INSTRUMENTAL CONDITIONING AND EXTINCTION

LANGDON E. LONGSTRETH[1]

University of Southern California

This monograph reports 3 experiments concerned with the concepts of secondary reinforcement and frustration as they apply to human instrumental conditioning. A review of the literature led to 2 conclusions: (a) there is little unequivocal evidence for secondary reinforcement at the human level, and (b) much of the data can be interpreted as supporting the extension of Amsel's frustration theory to human behavior. The present experiments explored these two preliminary conclusions. Experiments 1 and 2 paired 1 cue with reward (S+) and another with non-reward (S−), and then presented the cues alone, S+ to half the Ss and S− to the other half. S+ resulted in faster extinction as well as in other indications of greater frustration. There was no support for secondary reinforcement. Experiment 3 investigated reinforcement schedule and nearness to goal as they affected speed, amplitude, and resistance to extinction. The results once again failed to confirm secondary reinforcement predictions, and provided remarkable support for frustration theory.

THE present paper reports three studies of human instrumental learning and extinction. It is conceptually concerned with the empirical validity of two quite different concepts as they apply to human behavior: secondary reinforcement and frustration. Amsel (1961) and the author (Longstreth, 1964) have both drawn attention to possible confusions surrounding these two concepts and have wondered if one concept could not satisfactorily account for the data usually ascribed to both. The paper begins by examining the nature of this confusion.

The basic definition of a secondary reinforcer (Sr) is well known: it is a stimulus configuration which, through association with a reinforcer, acquires the capacity to influence preceding behavior in a way similar to that of the original reinforcer itself. In other words, responses followed by Sr are strengthened (learned) or at least maintained at a higher level than responses followed by neutral stimuli not previously paired with reinforcement.

A number of refinements have been suggested from time to time, each attempting to specify more clearly those operations which are *sufficient* for Sr effects. While these refinements are not central to the present discussion, three prominent ones may be noted in passing. Taken in chronological order, the first is a *discriminative stimulus* hypothesis formulated by Schoenfeld, Antonitus, and Bersh (1950). Attempting to rationalize their own failures to obtain Sr effects, they speculated that a stimulus must be a discriminative stimulus before it can function as an Sr. In view of its highly tentative nature (offered as a possible explanation of null results), it cannot be considered a highly developed conceptualization. Several subsequent studies have not supported it (e.g., Ratner, 1956; Wycoff, Sidowski, & Chambliss, 1958).

A more highly reasoned refinement is a *discrimination* hypothesis formulated by Bitterman and his associates (Bitterman, Feddersen, & Tyler, 1953; Elam, Tyler, & Bitterman, 1954). According to this position, a stimulus paired with reward during

[1] This research was supported by grants from the National Science Foundation (G16301) and the University of Southern California Faculty Research Fund. Experiments 2 and 3 were carried out while the author was a Joseph P. Kennedy Visiting Professor at George Peabody College.

acquisition and then presented alone in extinction serves to differentiate acquisition from extinction. To the extent the subject (*S*) discriminates this change, extinction will be faster. Thus the opposite of Sr effects are predicted under certain conditions, and two experiments confirm the prediction.

Finally, an *information* hypothesis has been suggested by Egger and Miller (1962). According to this notion, stimuli correlated with reward will become Sr's only if it is impossible to anticipate reward from other stimuli. In other words, if a stimulus is redundant in the sense that other stimuli have already "informed" *S* of impending reward, it should not acquire Sr properties. Although data supporting this hypothesis are reported by the authors, a more recent study fails to confirm it (McKeever & Forrin, 1966).

It is to be noted that all these elaborations share the two operations of pairing the to-be Sr with an established reinforcer and then presenting it alone. Of primary importance to the present discussion is the fact that a current theory of frustration specifies exactly the same operations as necessary for creating a *frustrating* situation. Reference is made here to Amsel's well-known frustration theory (Amsel, 1958, 1962). According to this position, frustration is defined as follows: "...*Frustrative events*—the absence of or delay of a rewarding event in a situation where it had been present previously [Amsel, 1958, p. 102]." According to this position, a frustrative event results in an unconditioned aversive emotional response (Rf), with much the same properties as fear. Just as components of the fear response can become attached to contiguous stimuli, thereby resulting in conditioned fear, so can components of Rf become conditioned, resulting in anticipatory (conditioned) frustration (rf). The implications of these assumptions have been verified in a number of investigations, and will not be reviewed here.

Let us now contrast the implications of Sr theory and frustration theory with a hypothetical example. Assume that rats are exposed to a successive discrimination problem: locomotion down a white (W) alley to a W goal box is reinforced, and down a black (B) alley to a B goal box is nonreinforced. After a discrimination has developed, extinction is introduced. Half the rats are run in W, half in B. Which group will extinguish first? According to the notion of secondary reinforcement (and note that the requirements of both the discriminative stimulus and information hypotheses are fulfilled by W), W is an Sr while B is not. Therefore *S*s run in W should be more resistant to extinction. According to frustration theory, exposure to the W goal box without reinforcement is frustrating, while exposure to the B goal box is not. The elicitation of Rf, and the subsequent elicitation of rf in the alley, will serve to inhibit the instrumental response in W, much as an animal learns to inhibit responses leading to a fear CS (i.e., passive avoidance conditioning). Thus Amsel's theory predicts *faster* extinction in W. The previously noted studies by Bitterman and his associates are very similar to this hypothetical study and, as noted, the results do not support Sr predictions. As is typical of all the animal studies concerned with secondary reinforcement, no mention is made of the possible role of frustration. Mowrer (1960), however, has pointed out the relevance of frustration theory to these studies, noting that it very nicely accounts for some of the results.

Both secondary reinforcement and frustration concepts have been applied to human behavior as well as to animal behavior. As with the latter, studies with human *S*s have been oriented towards *either* secondary reinforcement or frustration, but not both. Thus one line of studies has been cited as supporting Sr theory (a series of studies by Nancy and Jerome Myers are in this tradition, a recent one appearing in 1965), while another group, smaller but growing, is interpreted in terms of frustration theory (e.g., Haner & Brown, 1955; Holton, 1961; Longstreth, 1960, 1965; Ryan, 1965). There is no conceptual overlap between these two groups of studies: the existence of a second concept with the same operational definition but contrary implications is all but ignored. Perhaps this is because most of these studies were not designed to simultaneously evaluate both concepts, but only one. Thus failure to confirm the experimental hypothesis did

not automatically suggest the operation of an opposing concept.

The present studies were designed to clearly differentiate between Sr and frustration implications, so that the relative "truth value" of the two concepts could be determined. Thus conditions were sought which led to clearly incompatible predictions, but which always involved the basic operation of pairing a stimulus with reward and then presenting it alone. The first two studies are similar to the hypothetical rat study previously discussed, and the third study investigates goal gradient and partial reinforcement effects as they might be predicted from these two formulations.

Experiment 1: Discrimination Training Followed by Extinction to S+ or S−

Children were presented with a successive discrimination problem involving single, separate presentations of each of two visual stimuli (onset of lights varying in intensity). Turning off one of the lights resulted in a reinforcer (marble) while turning off the other light did not. Thus one of the lights was paired with reinforcement, as was the sound of the marble-ejection solenoid. A series of training trials was given, each trial consisting of presentation of one of the lights and its termination by *S* and subsequent delivery of a marble following the appropriate light. The instrumental response consisted of pushing a joystick to the left to turn off one light and to the right to turn off the other light. Extinction was then introduced (no marbles). One group was presented with both stimuli which previously had been paired with reward (the appropriate light before each response and the sound of the solenoid afterwards), a second with just one (the appropriate light, with the solenoid turned off), and a third with neither (presentation of the light previously *not* paired with marbles). Resistance to extinction was then determined, along with amplitude and latency measurements of each joystick response.

Method

Ss and apparatus. The *Ss* were 66 children from the second and third grades of the Manhattan Beach Public School System, Manhattan Beach, California. They were randomly assigned to three extinction groups of 22 children each, with 8, 10, and 11 males in the three groups.

The apparatus consisted of three main units: a stimulus-response unit, a control unit, and a recording unit. The first unit was situated on one side of a gymnasium and the other two units behind a curtain at the other end of the room (approximately 100 ft. away), so that the experimenter (*E*) could not only remain unobserved once the experiment started, but was also essentially incommunicado.

The stimulus-response unit consisted of a large black box 44 in. high and 34 in. wide. It rested on a low table so that a 9-in. square milk glass window mounted on the front surface was about even with *S*'s eyelevel. Directly below the window a joystick handle protruded which could be turned to the left or right. Springs returned it to a central position when pressure was released. To the right of the window and on the front was a clear plastic tube mounted in a vertical position. It received marbles which were automatically ejected into it on a programmed basis. The marbles rested on top of each other, thus filling up the tube as they accumulated. Its capacity was 40 marbles.

The control unit was simply a programming unit for stimulus and reinforcement events. The recording unit was a two-channel ink-flow Offner Dynograph. One channel was connected to a potentiometer whose resistance was determined by amount of displacement of the joystick, thus providing a permanent record of response characteristics. Marker pens signaled onset and offset of illumination of the milk glass window, as well as occurrence of marble delivery. It was thus possible to measure response amplitude and latency as well as number (frequency) of responses. Paper speed was 10 mm/sec, allowing accurate time measurements to the nearest tenth of a second. Amplitude was measured in millimeters of pen deflection from a base line.

Procedure

Teachers sent *Ss* to the experimental room one at a time. They were greeted by *E* and informed that he had a marble game for them to play and that if they earned enough marbles, these could be traded for a prize. Attention was drawn to the stimulus-response unit and *S* was seated on a low chair in front of it. The milk glass window, previously illuminated, was pointed out, and *S* was told the game consisted of turning it off whenever it came on by turning the response handle in the correct direction. The *E* demonstrated by turning off each of two illuminations twice. He then announced that "sometimes" a marble would be ejected into the plastic tube when *S* turned off the light and that when the marbles reached a marker on the tube *S* could trade them for a prize. The marker, a piece of tape, was placed high enough on the tube so that 20 marbles were required to reach it. The *S* was told he could stop whenever he wanted to, and following these instructions *E*

went behind the curtain and initiated the first training trial.

The experiment was programmed for two phases, training and extinction. During *training*, 36 trials were presented, involving 18 presentations of each of two illuminations in the stimulus window, a dim and bright illumination. Illumination was measured by a Weston Master IV exposure meter held 1 ft. from the stimulus window. Indirect illumination with the stimulus window off was 0.3 ftc. Onset of the dim illumination (D) resulted in a reading of 3.1 ftc., and onset of the bright (B) illumination resulted in a reading of 20 ftc. The lights, therefore, were easily discriminable. They were presented in a random order with the restriction that neither intensity appear more than three times in a row. A correct response terminated the illumination immediately, and the next intensity was automatically presented 4 sec. later.

The instrumental joystick response was simple to make, involving about a 30-degree arc in the correct direction to turn off the light. The amount of pressure required was minimal in order to reduce fatigue effects, but yet strong enough that the handle would return promptly to a central position when pressure was released. A pressure of 3 lb. moved it a sufficient distance in either direction. It could be rotated a maximum of about 90 degrees before a blocking device stopped further movement, thus making possible the measurement of variations in pressure. If *S* moved the handle in the wrong direction, he was allowed to correct his error before proceeding to the next trial. The *Ss* typically made only one or two errors at the very beginning of training.

For all *Ss*, termination of one light (S+) always resulted in automatic ejection of a marble 2 sec. later. Termination of the other light (S−) never resulted in a marble. Within each subsequent extinction group, direction of the correct response was counterbalanced with respect to light intensity, and light intensity was counterbalanced with respect to marble ejection.

Following the 36 training trials, *extinction* was introduced with no interruption or warning. Two groups were presented only with S+, but a correct response was no longer followed by marble ejection. One of these groups (S+n) was exposed to the noise of the marble solenoid 2 sec. after a correct response, just as during training. This mechanism was turned off for the other group (S+). The third group (S−) was presented with only the negative illumination.

The experiment was terminated when *S* said he wanted to stop (almost invariably accompanied by rising from his chair) or when a 10-sec. time interval intervened between light onset and a response. If neither criterion had been reached by 150 extinction trials, *E* terminated the experiment. In any event, *S* was told it was uncertain whether he had won a prize or not, and he would have to wait until everybody had played the game before the "winners" could be determined. At the termination of the study, every *S* was given his choice of several prizes.

RESULTS AND DISCUSSION

Light intensity and response direction per se exerted no significant effects on response strength, so that collapsing of these counterbalancing subgroups was justified. Amplitude measurements, however, were affected by direction of the preceding response; a second response in the same direction was weaker than a response in the opposite direction. In order to remove this source of variation from the training data, trials were chosen on which the preceding response was in the same direction. Pairs of such trials were then combined to smooth the resulting learning curves. The ordinal values of these trials varied for various counterbalancing subgroups, but were constant across the three extinction conditions. Three data ranges were picked in order to present performance near the beginning, middle, and end of training. Trials 2, 3, 4, and 8 were used to compute the first curve point, Trials 13–16 for the second, and Trials 31, 34, 35, and 36 for the last point.

The first and third vertical panels of Figures 1 and 2 present amplitude and speed data (reciprocal of response latency), respectively, for these training trials, separately for each extinction condition as well as for all three conditions combined. Figure 1 indicates a gradual divergence in amplitude over training trials, with amplitude to S+ increasing relative to S−. The reliability of this trend was evaluated by a mixed three-way analysis of variance (Lindquist Type VI) with extinction condition as a between-*S* variable and trials and stimuli (S+ versus S−) as within-*S* variables. The only F ratio approaching significance was the trials-by-stimuli interaction, with an F of 7.94 ($df = 2/126, p < .001$). It may thus be concluded that amplitude to S+ increased relative to S− amplitude over training trials, and that all three subsequent extinction groups manifested the same trend.

Speed data (Figure 2) present a similar picture except that the S+, S− divergence is not so pronounced. Analysis of variance indicated that while speeds to S+ were significantly faster than speeds to S− ($p <$ .001), the divergence was not significant. The first trial of the pair of trials deter-

mining the first curve point was then employed by itself to determine the first point, under the hypothesis that learning was extremely rapid in this situation and that data more representative of the *beginning* of learning would more likely reveal a divergence if any, in fact, existed. Single-trial data were also employed to determine the other two curve points in order that distributions with comparable variance be provided at all three curve points. An inspection of the resulting curves revealed an increased amount of divergence from that depicted in Figure 2 for all three groups, and statistical analysis confirmed the reliability of this trend by yielding a significant trials-by-stimuli interaction ($p < .05$). It may thus be concluded that the marble was indeed a reinforcer: responses followed by it developed greater strength and speed than responses not followed by it.

The second and fourth vertical panels of Figures 1 and 2 present amplitude and speed measurements during extinction. Amplitude is presented for the second extinction trial and for successive fifths of extinction. The first extinction trial is not represented because it involved a change in response direction from the preceding trial for half the Ss. Speed data are presented only for the first six extinction trials because the event-marker pens were turned off at that point, making further speed measurements impossible.

Amplitude data reveal an approximately linear decrease over extinction trials, with about the same slope for all three conditions, and with greater amplitude to S+ than to S−. Analysis of variance indicated only the trials effect to be significant ($p < .001$). Speed data reveal a large drop from the first to the second extinction trial for Condition S+n, a smaller drop for Condition S+, and no drop at all for Condition S−. The decrements for Conditions S+n and S+ were significant as measured by related *t* tests ($p < .05$). Variances of the difference scores between the first two extinction trials

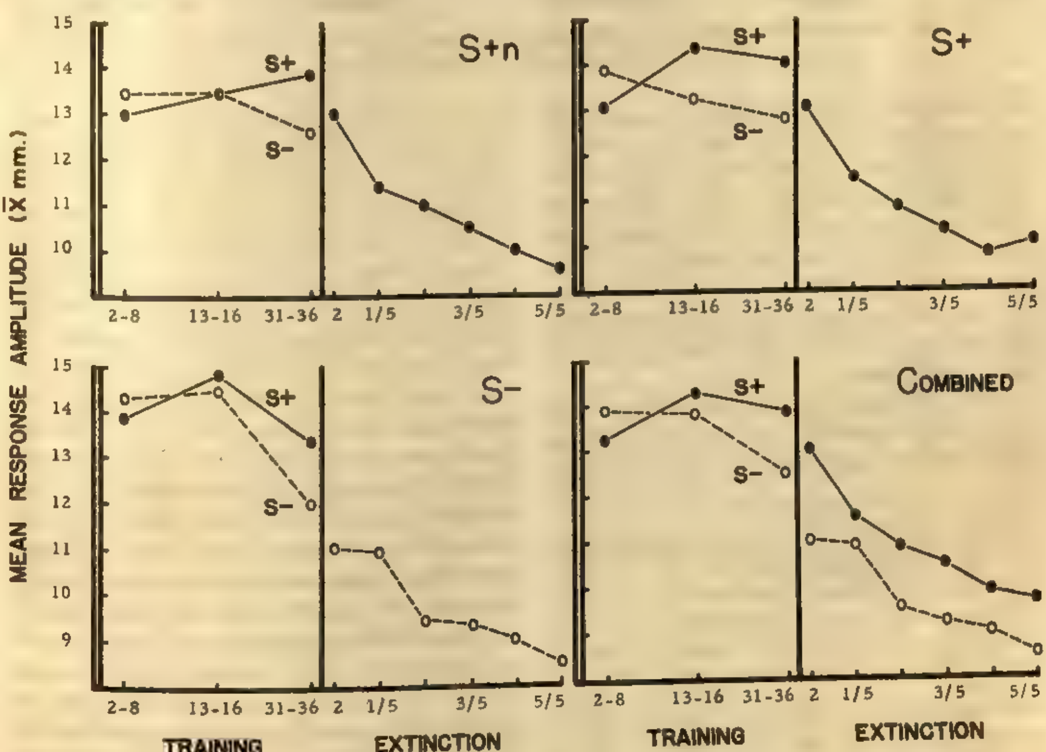

FIG. 1. Mean response amplitude in training and extinction for the three extinction conditions separately and for all conditions combined.

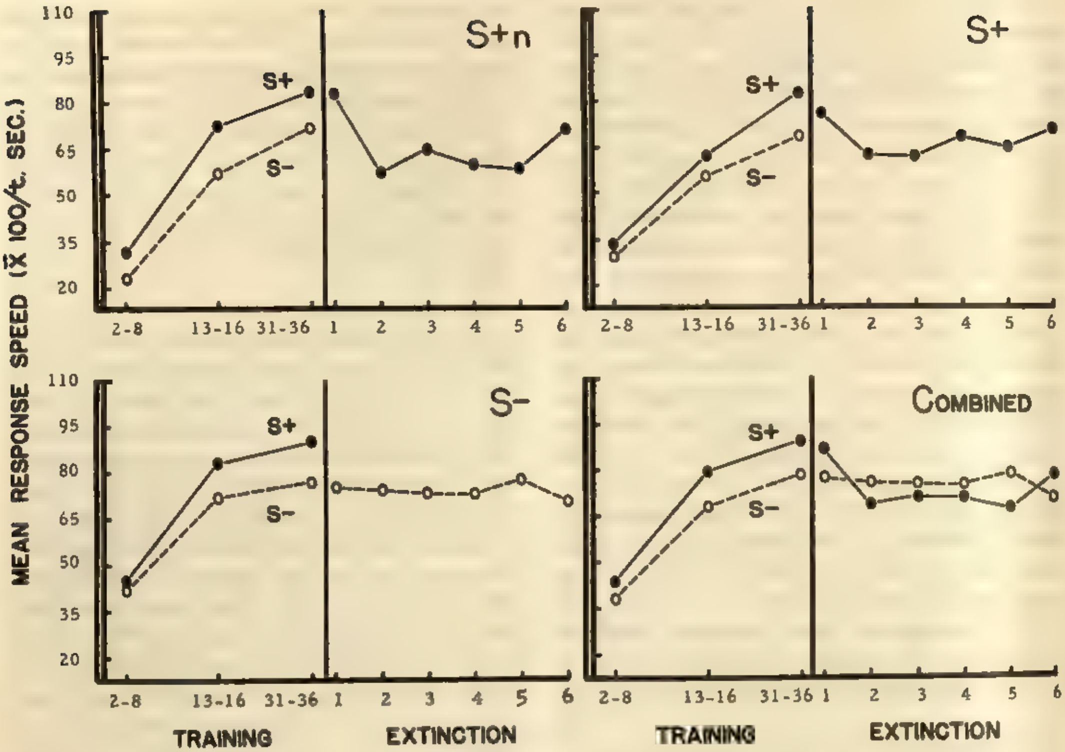

FIG. 2. Mean response speed in training and extinction for the three extinction conditions separately and for all conditions combined.

were also different, being 398, 716, and 246 for Conditions S+n, S+, and S−, respectively ($p < .05$, Hartley test). These differences were maintained through Extinction Trial 6, at which point speed variances were 1,866, 752, and 452 for Conditions S+n, S+, and S− (F max. $= 4.13$, $p < .01$). Differences in mean speed for Extinction Trials 2–6 were not significantly different for the three conditions.

Table 1 presents the number of *S*s in each group who responded less than 50, 100, or 150 times during extinction, as well as the number who responded more than 150 times. All *S*s but one of those who extinguished prior to 150 responses met the criterion of verbalizing a wish to stop, while the one additional *S* met the criterion of a 10-sec. response latency. These frequencies clearly indicate least resistance to extinction for Condition S+n and greatest resistance to extinction for Condition S−. Evaluation of this trend was carried out by dividing each group into those above and below the approximate overall median (150) and computing chi-square. A value of 13.56 was obtained ($df = 2$, $p < .005$). Differences between S+n and each of the other conditions were also significant, while the difference between S+ and S− was not significant.

TABLE 1

DISTRIBUTION OF *S*s FOR NUMBER OF RESPONSES TO EXTINCTION (n)

n	Condition		
	S+n	S+	S−
150+	4	12	16
100–149	2	2	1
50–99	6	2	2
0–49	10	6	3

These findings do not support predictions based upon secondary reinforcement theory, which predicts the *opposite* ordering of groups in terms of resistance to extinction. Considering frustration theory, the following interpretation is patterned closely after

Amsel's use of the theory with infrahuman data.

Condition S+n maximized the similarity of the extinction condition to the reinforced training condition, and thereby elicited the strongest reward anticipations (indeed, a number of *S*s were observed to look at the marble ejection tube when the delivery solenoid was activated). Not receiving the reward, this condition elicited maximum frustration responses. The aversive nature of these responses elicited avoidant response tendencies which then conflicted with subsequent instrumental response tendencies. If the instrumental response then occurred, however, it should have been "amplified" by the drive properties of frustration. Eventually, continued exposure to nonreinforcement strengthened the avoidance response tendencies above those of the instrumental response, and extinction occurred. The same processes operated to a lesser extent in Condition S+, and least in Condition S−.

How well do the facts fit the theory? The extinction data, of course, are in the predicted order, although the difference between Conditions S+ and S− is not significant (Experiment 2 follows up this matter). The predicted conflict in Condition S+n (and, to a lesser extent, in Condition S+) is also supported in terms of speed means and speed variability. That is, speed scores are conventionally used to infer conflict (i.e., Castaneda & Worrel, 1961; Finger, 1941), the assumption being that strong conflict results in temporary response blockage, and hence in long latencies. Condition S+n produced the greatest drop in starting speed following the first extinction trial, and Condition S− the least, thus confirming the prediction. Speed variability of early extinction trials was also greater for Conditions S+n and S+. Amsel has interpreted such variability to be indicative of the waxing and waning of the conflicting response tendencies, with momentary dominance of the instrumental response producing a short-latency response, and momentary dominance of the avoidance response tendency producing a long-latency response. Thus speed variances as well as speed means conform to expectations.

The heightened drive prediction, however, was not confirmed: response amplitude following the first extinction trial was not affected by extinction conditions, and hence was not a function of frustration. Experiment 2 also investigated this matter in greater detail.

Experiment 2: A Pseudoreplication of Experiment 1

The most impressive finding of Experiment 1 was the rapid extinction under Condition S+n, the condition providing both stimuli previously paired with marble reinforcement. Although this finding contradicts secondary reinforcement theory and supports Amsel's frustration theory, the lack of a significant difference between Conditions S+ and S− is disturbing to both positions. Experiment 2 had as one goal a second estimate of the population difference between these two conditions. It thus consisted of two groups, S+ and S−. A pretraining modification was suggested by the possibility that light offset itself was reinforcing and hence masked Sr effects attributable to the marble. Although the greater amplitude and speed scores to stimulus S+ during training argues against this possibility, it was nevertheless decided to reduce these possible reinforcing effects by "satiating" *S* prior to the experiment proper. To this end a pretraining phase was added, consisting of 120 trials (60 B and 60 D), *S* terminating each light by the appropriate joystick response. It was assumed that the novelty of the "game" would be reduced substantially by this exposure.

Two other modifications had to do with the failure to find increased response amplitude following initial extinction trials under Conditions S+n and S+, that is, failure to observe Amsel's FE. First, a number of *S*s were responding with maximum amplitude *prior* to extinction, since very little effort was required to turn the handle as far as it would go. For these *S*s, a ceiling effect was therefore operating, making it impossible for amplitude to increase with the onset of extinction (separate analyses of *S*s not responding at maximum still failed to show the FE, but the *N* was quite small). Tension on the response handle was therefore increased to the point where it was rather

difficult to turn it a maximum distance. Second, it is possible that the FE was minimized by the temporal interval between trials (4 sec.). Perhaps the postulated aversive motivation produced by nonreinforcement dissipated prior to the next response. The duration of the intertrial interval was thus introduced as a second independent variable in Experiment 2, retaining a value of 4 sec. for half the *S*s and assuming a value of 2 sec. for remaining *S*s.

Two final modifications were introduced. First, a more complete picture of extinction behavior was obtained by prolonging its possible duration and securing speed as well as amplitude measurements on all extinction responses. Second, *S*s were drawn from a population of institutionalized mental retardates rather than from a public school (see Footnote 1).

METHOD

*S*s *and Apparatus*

Thirty-two mental retardates served as *S*s. All were from Clover Bottom Hospital and School, Doneldson, Tennessee, a state institution for the mentally retarded. They were assigned in prearranged order to one of eight conditions, the second 16 *S*s being assigned in the same order as the first 16. Sex, CA, and MA were ignored, except that *S*s so profoundly retarded that they could not follow instructions were not used. Mean CA for the two conditions, S+ and S−, was 21.6 and 21.1 yr., respectively, while mean MA was 6.6 and 7.4 yr., respectively. Neither of these differences approached significance. The number of females in conditions S+ and S− were eight and four, respectively. Since there was a nonsignificant tendency for females to extinguish more slowly than males, and since it was found that Condition S+ resulted in *faster* extinction, sex differences were ignored.

The apparatus was the same as used in Experiment 1. The stimulus-response unit was placed in a room adjacent to *E*'s room, the two being separated by a partition holding a one-way window. The control and recording units were in *E*'s room and were connected to the stimulus-response unit by means of an electrical conduit. The springs used to maintain tension on the response handle were replaced so that about 10 lb. pressure in either direction was required to rotate the handle its maximum distance.

Procedure

An assistant brought *S*s to the research building one at a time. After *S* was informed that there was a game for him to play, he was shown the stimulus-response unit and told that the game consisted of turning off the light by turning the joystick in the correct direction. After two demonstrations with each light by *E*, the pretraining phase was instituted, consisting of 60 presentations each of B and D in a mixed order, the second sequence of 60 trials being identical to the first sequence. Except for three "non-offset" trials, a correct response terminated the light immediately and the next light was automatically presented 2 or 4 seconds later, depending upon the intertrial interval condition to which *S* belonged. All *S*s but one were responding correctly by the end of pretraining, and that *S* was replaced.

On Trials 70, 80, and 100 a correct response did not result in immediate offset of the light. Instead, the response had to be repeated two more times before the light terminated. The purpose of these trials was simply to investigate the effects of this kind of treatment.

Following pretraining *E* entered the room and announced that he had a new game. The *S* was told to continue turning off the lights, but that "sometimes" a marble would fall down the tube into a cardboard box placed directly below the tube. At this point *E* produced a steel container completely full of 30 steel marbles. After commenting on how full the container was and making sure that *S* saw it, *E* put the marbles, one by one, into the top of the apparatus and out of sight. Then he placed the empty steel container beside the box at the end of the tube and told *S* to put the marbles back into the container whenever one fell into the box. He was told that if he could fill the container with marbles "just like it was before," he could win a prize. At this point a nickel and a 5-cent bag of M & M candy were produced and *S* was asked which he would rather win. His choice was placed next to the empty steel container. After telling him he could quit whenever he wanted to by getting up and coming into the next room where *E* was waiting, *E* returned to the control room and initiated the training phase.

The preceding change in incentive conditions from Experiment 1 was to insure that these retarded *S*s would notice the pairing of S+ and marbles; thus they were forced to handle each marble instead of *perhaps* looking at it in the tube.

Training consisted of 18 presentations of each light with the same mixed sequence for all *S*s, just as in Experiment 1. *S*s in the 4-sec. intertrial interval condition received the marble two sec. after responding to S+, and *S*s in the 2-sec. condition received the marble 1 sec. after a response to S+.

Extinction followed training with no interruption. Half the *S*s were exposed only to S−, and the other half were exposed to the following sequence: S+, S+, S−, S−, S+, etc., that is, two presentations of each stimulus followed only by S+. If *S* had not actually walked out of the room by 54 extinction trials, *E* entered the room and said, "Do you want to quit or play some more?" If there were any questions *E* said, "You can play as long as you like." If *S* indicated he wanted to continue, *E* returned to the control room. This

process was repeated at the end of 154 trials if *S* had not extinguished. If *S* did not quit at that point, he was allowed to continue uninterrupted until 750 trials had occurred ($N = 2$) or until the next *S* arrived ($N = 2$). The artificial stopping of these four *Ss* was, of course, taken into account in the analysis. The purpose of the interruptions demands some explanation. It was felt that making it "easy" to stop served two purposes. First, it counteracted any tendency these *Ss* may have learned in the institution to obey orders, a tendency which seemed predominant at this institution. Second, it served to reduce the possibility that *S* would either not notice or forget the fact that it was acceptable behavior to stop responding when he wished.

The three factorially combined variables, then, were light-marble condition (B+ or D+), intertrial interval, and reinforcement interval (4 and 2 sec., respectively, for one condition, and 2 and 1 sec., respectively, for the other condition), and extinction condition (S+ or S−). There were four *Ss* in each of the eight cells.

RESULTS

The overall median number of responses to extinction for all 32 *Ss* was 60. The medians for extinction Conditions S+ and S− were 38 and 117, respectively, yielding the contingency relationship described in Table 2. Chi-square, corrected for noncontinuity, equaled 4.05, $df = 1$, $p < .05$, indicating faster extinction under Condition S+. The four *Ss* stopped artificially did not affect these values since they all responded more than 400 times (three were in Condition S−). The direction of the difference observed in Experiment 1 between these two conditions was therefore replicated, this time reliably.

Experiment 2, like Experiment 1, thus did not yield Sr effects, but rather the opposite. While these results are consistent with frustration theory, more demanding evaluations can be made by examining amplitude and speed data, as in Experiment 1. Compared

TABLE 2

NUMBER OF *Ss* IN CONDITIONS S+ AND S− ABOVE AND BELOW THE GENERAL MEDIAN NUMBER OF RESPONSES TO EXTINCTION

Median number of responses to extinction	Condition	
	S+	S−
60+	13	3
59−	3	13

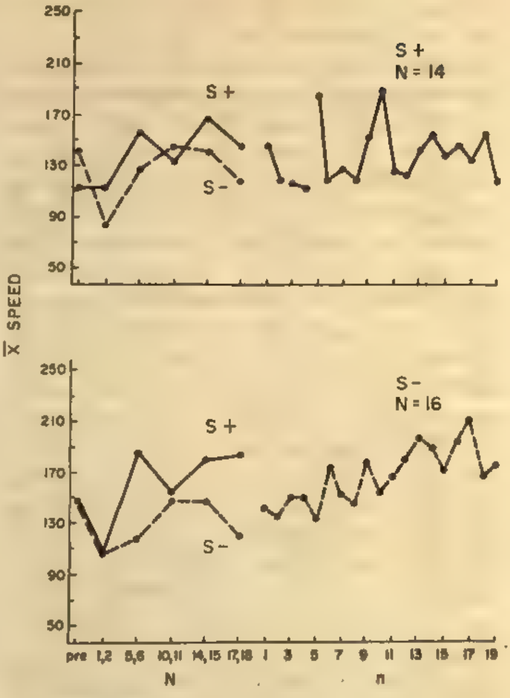

FIG. 3. Mean response speed (100/t sec.) in training (N) and extinction (n) for the two extinction conditions, S+ and S−.

to S−, the S+ condition would be expected to (*a*) produce a greater decrement in starting speeds after the first extinction trial; (*b*) produce greater speed variability on early extinction trials; and (*c*) produce an *increment* in response amplitude, followed by a decrement on later extinction trials.

Figure 3 presents mean response speeds during training and the first 19 extinction trials, and Figure 4 presents amplitude data.[2] The intertrial interval (ITI) conditions of 2 and 4 sec. are combined in these figures, even though the 2-sec. interval resulted in consistently greater frustration effects as subsequently described. The training data are similar to those of Experiment 1: both speed and amplitude curves indicate discrimination between S+ and S−, with

[2] Extinction data in Figures 3 and 4 are limited to 19 trials because the first *S* in Condition S− extinguished at that point. Two *Ss* in Condition S+ who extinguished prior to 19 trials are excluded in order to preserve the continuity from training to extinction. The nature of the curves is not altered by exclusion of these data.

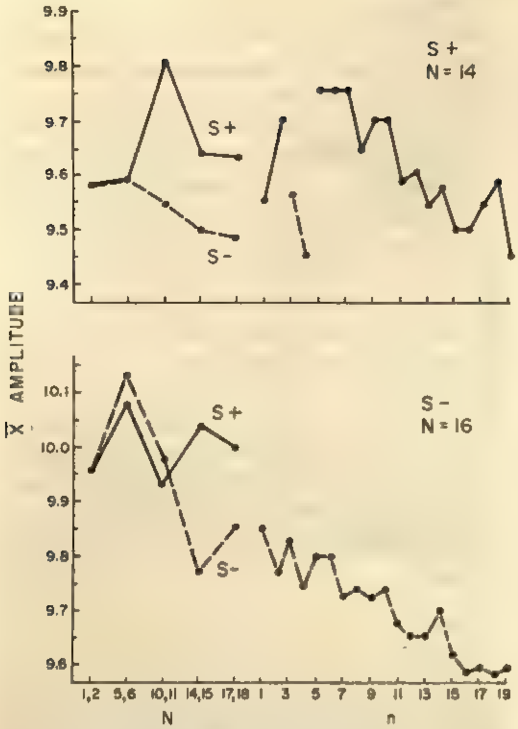

FIG. 4. Mean response amplitude (in millimeters) in training (N) and extinction (n) for the two extinction conditions, S+ and S−.

greater speeds and amplitude to S+. The divergence in the speed curves was not significant unless pretraining trials were included (represented by "pre" in Figure 3), confirming the results of Experiment 1 in suggesting that speed is a more sensitive index of discrimination than amplitude in the present situation.

Turning next to response speeds in extinction, Figure 3 indicates that the onset of extinction resulted in a greater decrement in Condition S+ than in S−, as predicted. This trend was evaluated by calculating speed differences between the first two extinction trials and performing a t test on the two resulting means. The difference of these differences was significant ($p < .05$), providing statistical support for the prediction and confirming the results of Experiment 1.

An unexpected finding is the gradual increase in speed during extinction for Condition S−. This increase from the end of training to Extinction Trial 19 is significantly greater than the corresponding change in Condition S+ ($p < .05$). Indeed, by the nineteenth extinction trial *Ss* in Condition S− were responding with a speed equal to that elicited by the S+ stimulus at the end of training. This pattern persisted throughout extinction, in spite of the greater fatigue effects under Condition S−, where a greater number of responses occurred.

Speed variability may be considered next. As the irregular extinction curve in the top half of Figure 3 suggests, Condition S+ produced greater variability than Condition S−. Evaluation of this difference necessitated the matching of Conditions S+ and S− as closely as possible on the basis of mean speeds, due to a positive correlation between means and variances. Such matching was possible on three extinction trials: 6, 9, and 18. The means and variances for the two conditions on these trials are presented in Table 3. In spite of the fact that mean speeds are a little larger for Condition S−, thus biasing variances in the direction of greater S− variability, the table shows larger variances for Condition S+ in all three comparisons, and in two the difference is significant. Speed variances, therefore, conform to expectations, suggesting greater conflict in Condition S+.

Finally, response amplitude may be examined. Figure 4 shows that amplitude during early extinction was considerably different for the two conditions, with S+ showing a large increment from the first to the second extinction trial, and S− a slight decrement. The S+ increment is significant as measured by related t, $t = 2.41$, $df = 13$, $p < .05$, while the decrement for S− is not

TABLE 3

SPEED MEANS ($\bar{X}$) AND VARIANCES (s^2) ON EXTINCTION TRIALS (n) FOR CONDITIONS S+ AND S−

n	Condition S+ $\bar{x}$	Condition S+ s^2	Condition S− $\bar{x}$	Condition S− s^2
6	131	8,195	136	2,783*
9	141	13,541	146	2,894*
18	155	19,775	164	11,803

* $p < .05$.

significant. The third and fourth extinction trials for Condition S+ consisted of two presentations of S− and are represented by a dotted line. The first presentation of S− indicates an increment in amplitude from the end of training, while a decrement is indicated for the third extinction trial in Condition S−. These differences between the two groups are significant: when *S*s were matched on the basis of amplitude to S− at the end of training, and using all 16 *S*s in Condition S+, Condition S+ resulted in a larger increment from the end of training to Extinction Trial 3 than Condition S−, related $t = 2.11$, $df = 15$, $p < .05$. This statistic leads to the conclusion that two presentations of S+ without reward led to a greater increment in amplitude to S− than two preceding presentations of S−. The importance of this conclusion is discussed shortly.

Amplitude from Extinction Trial 5 shows a further increment for three trials under Condition S+, and then a gradual decline, whereas a monotonic decrement throughout extinction is depicted in Condition S−. It may be concluded, therefore, that response amplitude in Condition S+ is distributed across extinction trials as an inverse-*U* function, as predicted. When the data of the 2- and 4-sec. ITI conditions are examined separately, the inverse-*U* function appears in both conditions, but is larger in the 2-sec. condition. It may thus tentatively be concluded that part of the reason it was not observed in Experiment 1 was due to the intertrial interval.

Discussion

Several additional comments about Experiment 2 need to be made. First, although increased amplitude following nonreward is often interpreted as indicative of the nondirective motivational properties of frustration, an associative explanation (e.g., Brown, 1961) is not ruled out. That is, it could be argued that *S* preexperimentally learned to *try harder* following failure, and hence the increased amplitude is a habit phenomenon rather than a motivational one. If such were the case, one might not expect an increase in amplitude to S−, since the response to this cue never resulted in "failure." Yet an increase was observed on Extinction Trial 3, where S− was presented following two previous nonreinforced occurrences of S+. Furthermore, this increase was about twice as great with a 2-sec. intertrial interval as with a 4-sec. interval. Both findings are more consistent with a motivational interpretation than with an associative interpretation.

Second, the decrement in speed under Condition S+ may not necessarily have been due to frustration-mediated avoidance tendencies. It was observed that *S*s often oriented their faces toward the cardboard marble container after responding to S+, and maintained this orientation until the marble was delivered. In extinction, this posture was often maintained until the next presentation of S+, at which time *S*s again looked at the stimulus window and made their response. Thus they often were not looking at the window when the stimulus was presented. Perhaps, then, the decrement in speed was the result of orienting responses which interfered with performance of the joystick response. Fortunately, it was possible to test this argument. It will be recalled that a different type of "nonreinforcement" was encountered on three pretraining trials: the stimulus light did not immediately terminate with a correct response. The important point is that the locus of "nonreinforcement" on these trials was identical to the locus of stimulus presentation; that is, the stimulus window. Hence orienting responses elicited by nonreinforcement would be expected to facilitate speed of the next response, rather than to interfere with it. However, such was not the case. Mean speeds of one trial preceding and one trial following the first nonoffset trial (Trials 69 and 71) were 143 and 110, respectively, yielding a related t of 2.09, $df = 24$, $p < .05$ (seven *S*s who made an error on one of the two trials were discarded). This decrement of 33 units is similar to the decrement from the end of training to the second extinction trial for Condition S+ (27 units) and negates an interpretation based on changes in orienting responses.

Third, and last, the gradual increase in

speed to S− during extinction is puzzling but not entirely baffling. It may represent an increasingly strong expectation for S+, since *S*s presumably learned during pretraining and training that S− was always followed sooner or later by S+. It is significant that speed showed a final decrement on the last two-fifths of extinction, suggesting that these expectations finally weakened, contributing to extinction.

Experiment 3: The Effects of Reinforcement Schedule and Extinction Block Point on Instrumental Behavior

Experiments 1 and 2, taken as a whole, provide strong support for Amsel's frustration theory as it might be applied to human behavior. As far as secondary reinforcement is concerned, one may conclude that at best it was not demonstrated and at worst it was disconfirmed. But perhaps the paradigm and specific procedures used in these studies produced idiosyncratic results. It was therefore decided to approach the problem from a different point of view. The strategy was the same: to consider implications of frustration and secondary reinforcement theory and to oppose them, if possible, in a single experiment.

Amsel has used his theory to explain both training and extinction data obtained from a partial reinforcement schedule. The pertinent facts are these: in training, partial reinforcement results in (*a*) initially slower speed, (*b*) final higher speeds (if more than 30–40 trials are given), and during extinction in (*c*) greater resistance to extinction. There is some evidence that (*a*) and (*b*) are dependent upon where measurements are taken in the response sequence; speeds recorded in the early or middle parts of the locomotor sequence confirm these findings, while measurements at the end of the sequence, for example, in the goal area, suggest faster speeds for 100% reinforcement throughout training; that is, such measurements support (*a*) but not (*b*) (Amsel, 1964; Wagner, 1961). It was decided to test Amsel's formulation at the human level by attempting to reproduce these findings. The experimental design thus included two main conditions, a 100% and a partial reinforcement schedule. In order to investigate the dependency of (*b*) upon the particular response segment measured, the instrumental sequence was divided into four discrete parts, with amplitude and speed measured independently for each one. More specifically, a trial consisted of successive presentations of four light intensities, progressing from B to D for half the *S*s and from D to B for the other half. Just as a rat "terminates" each segment of the alley with the same locomotor response, so was each light intensity terminated with the same joystick response: a movement to the left. A marble followed the last response, and then the first light in the sequence was presented again. Half the *S*s obtained a marble after every sequence, and half after only some sequences, as subsequently described.

Secondary reinforcement implications in a sequential task were first investigated at the human level by Lambert, Lambert, & Watson, (1953). Children learned to turn a crank and to insert tokens in order to earn candy at the end of the instrumental sequence. They were then extinguished at different "block points" from the "goal"; some *S*s earned tokens and were exposed to other stimuli in the sequence while other *S*s did not; they turned the crank but progressed no further. The reasoning was that *S*s extinguished near to the goal would be exposed to secondary reinforcement from the cues at the end of the sequence, and thus would be more resistant to extinction than *S*s extinguished further from the goal. The results, unfortunately, were exactly the opposite, and dramatically so: *S*s extinguished near the end of the sequence extinguished significantly *faster*, and with practically no overlap with *S*s extinguished near the beginning of the sequence. A second study, with some changes to control for an "uncontrolled" variable, found opposite results, but the differences were not nearly so dramatic.

It was decided to repeat the essential features of the Lambert et al. study, since such data are obviously relevant to both frustration and secondary reinforcement concepts. Therefore one extinction condition (1) involved presentation of only the

first light in the sequence (S1), and this light was presented over and over until *S* extinguished. A second condition (4) involved presentation of only the last light in the sequence (S4), and a third condition (1234) involved presentation of the entire sequence (S1–S4) until *S* extinguished. A frustration point of view would predict greatest frustration for Conditions 1234 and 4, since these conditions are more similar to the reinforcement situation than is Condition 1. Contrary to Sr predictions, these conditions would thus be expected to produce faster extinction and to produce other indications of frustration as well: temporary increases in amplitude, decrements in speed, and greater speed variability.

The experimental design was thus a 2 × 2 × 3 factorial one, consisting of reinforcement schedule (100% or partial), light sequence (B to D or D to B), and extinction block point Conditions 1, 4, or 1234.

Method

Ss and Apparatus

Sixty mental retardates from Clover Bottom Hospital served as *Ss*. None had been exposed to the apparatus before. They were exposed to 1 of the 12 conditions of the experiment in the following way: the first 2 served in Condition 1; the next 2 in Condition 2; etc.; until four *Ss* had served in each condition. Twelve remaining *Ss* were then assigned 1 to each condition, yielding a total of 60 *Ss*. Eight others were rejected and replaced, four because of apparatus failure, two because of extreme slowness in motor behavior, one because of epilepsy, and one because of poor eyesight.

As in Experiment 2, subject factors were ignored in the selection of these *Ss*. Table 4 presents means and *SD*s of CA, MA, and IQ for the six major conditions of the experiment (counterbalancing of light intensity and response sequence is not included), as well as sex distribution. It indicates that MA and IQ were highly similar for all six conditions, and that CA was similar for five of the conditions, but was only 19.2 for the sixth condition. Analysis of variance, however, indicated that CA differences were not significant for either experimental variable depicted in Table 4, nor for their interaction.

As a final check on the possible role of MA, four *Ss* were selected from each of the six cells in Table 4, one pair representing the highest MAs and the other pair the lowest MAs in that condition. Two groups of 12 were thus formed, representing the extremes in MA, with averages of 5.2 and 8.6 years. Mean number of responses to extinction for these two groups were 174 and 178, respectively, a difference too small to merit statistical evaluation (the ranges were 12–750 and 6–750).

Sex was not evenly distributed in the six conditions, although marginal totals reveal that it was evenly distributed for either experimental variable. The mean number of responses to extinction for the 31 males was 200, and the 29 females, 212, a difference not approaching significance.

The apparatus was exactly the same as used in Experiment 2, except that a modification in the incentive conditions was instituted. These conditions are subsequently described.

Procedure

As *E* and *S* entered the experimental room, *E* said he had a present for *S*. Inside the room a large, heavy Christmas package was sitting on a chair, attractively wrapped in gay colors and tied with a pretty ribbon. *E* handed it to *S*, telling him it was his. As *S* took it (and thus felt how

TABLE 4

Means ($\bar{X}$) and Standard Deviations (*SD*) of CA, MA, and IQ, and Sex Distribution, for Reinforcement Schedule (% Reinforcement) and Extinction Block Point

Percentage of reinforcement	Extinction block point								
	1234			4			1		
	CA	MA	IQ	CA	MA	IQ	CA	MA	IQ
100%									
$\bar{X}$	26.4	7.0	45.6	23.0	6.0	42.3	25.8	6.5	42.9
SD	7.1	1.9	7.3	7.3	0.9	7.9	7.9	0.8	7.3
Sex		7M,3F			6M,4F			3M,7F	
54%									
$\bar{X}$	23.7	7.1	46.8	19.2	6.7	46.7	25.2	6.4	43.2
SD	8.1	1.4	7.7	7.7	1.9	10.5	10.4	1.6	10.9
Sex		4M,6F			6M,4F			5M,5F	

heavy it was), *E* pulled it away and added, "Before you can open it, I have something for you to do." An 8-ft. board was sitting on two chairs beside the stimulus-response unit. It had 63 holes countersunk in a wavy line to form retainers for marbles, and with a colored line leading from one hole to the next, and with two lines after the last hole forming the outline of an arrowhead. The present was placed at the end of the board, so that the arrowhead pointed directly at it. As *E* placed the present on the board he said, "You have to fill all these holes with marbles, way down here to the very last one, before you can have the present. Let's put it here behind the last hole. Now, see this light here in the window? (The stimulus window had been previously turned on.) Well, you have to turn it off whenever it comes on; as soon as it comes on you turn it off. Then wait for it to come on again and then turn it off again. You turn it off by pushing this lever here. Watch. (*E* turned the lever to the left, turning off the light; when it automatically came on 2 sec. later, he turned it off again.) Now I'm going to tell you something else. *Sometimes* when you turn off the light, a marble will pop out of this tube and land in this box here. When it does, pick it up and put it in the next hole, and when you have them all filled up, you can have your present. Now you turn off the light to see if you can do it. (After *S*'s second response a marble was automatically ejected. The *E* uttered an exclamation of delight, told *S* to pick it up and place it in the first hole, which *S* did.) Now, when you are all finished, or whenever you want to quit, you just get up and come out the door to my room down the hall and tell me—I'll be doing some work in there. When you want to quit, come and tell me. Go ahead and begin now."

The *E* then remained in the room with *S*, paraphrasing and repeating the above instructions until it was obvious that *S* understood the nature of the task. He remained in the room until *S* was responding with less than maximum pressure on the joystick, but under no conditions remained after the first 30 trials: all responses after the thirtieth trial were performed alone. If *S* was responding habitually with maximum pressure, thus allowing no margin for variation in amplitude, *E* said, "You do not have to push that hard; it is easier if you push softly," and demonstrated by putting his hand over *S*'s on the response handle and turning the handle softly to the left. When *E* later returned to the control room he marked the recording paper with a pencil to indicate the specific trials he was in the room with *S*. Sometimes he went back two or three times within the first 30 trials to remind *S* once again that he did not have to "push so hard." It is to be noted that these instructions would tend to minimize the FE, and thus operated against frustration predictions.

Training. Training consisted of 61 trials of the S1–S4 sequence. Half the *S*s received a marble at the end of each sequence (100% reinforcement) and half at the end of 33 of the sequences, defining the partial reinforcement condition (54%). The *S*s in the latter condition found 28 marbles in the board at the beginning of the experiment. Thus all *S*s possessed 61 marbles at the end of training and needed 2 more to fill the board. The reinforcement schedule for the 54% condition consisted of four cycles of a 15-trial pattern and one final reinforcement. The reinforcement schedule for each 15-trial pattern was +, −, +, −, −, −, +, +, −, +, +, +, −, −, +. Each light came on 2 sec. after termination of the preceding light, and the marble was ejected 1 sec. after termination of S4. The interval between S4 and the subsequent onset of S1 was also 2 sec. Thus the interstimulus and intertrial intervals were both 2 sec.

Extinction. Extinction was introduced with no interruption on Trial 62. One-third of the *S*s from the 100% reinforcement condition were presented with only S1, another third only with S4, and the last third with the entire sequence S1–S4. A similar procedure was followed for 54% *S*s, so that 10 *S*s were in each of the six reinforcement schedule-extinction conditions. The extinction criteria were more stringent than in previous experiments; *S* was required either to leave the room or to hesitate with a 60-sec. latency before the experiment was terminated. It was thus possible to employ a number of different extinction criteria up to a limit of 60 sec. in order to determine their interrelationships. If *S* had not extinguished by 116 or 294 trials, *E* entered the room and said, "You can quit if you want to. Which would you rather do, quit or play some more?" If *S* did not want to quit *E* left the room. He did not enter it again after the two hundred and ninety-fourth response.

RESULTS

Training Amplitude

Extinction conditions were combined in order to provide a more stable picture of training data as a function of reinforcement schedule. The first 30 trials were not analyzed because of *E*'s interaction with *S* on some of these trials. For Trials 31–61, trials were selected on which the preceding trial had been reinforced for all *S*s. Trials following reinforcement were grouped into triads and the median amplitude recorded for each triad, separately for each of the four stimuli within a trial. Four such triads were obtained, covering Trials 34 to 58. The means of the medians of these triads are presented in Figure 5 separately for the two reinforcement schedules. This figure indicates that amplitude across trials was quite stable, that it systematically varied within trials, and that it varied as a function of reinforcement schedule. Responses to the

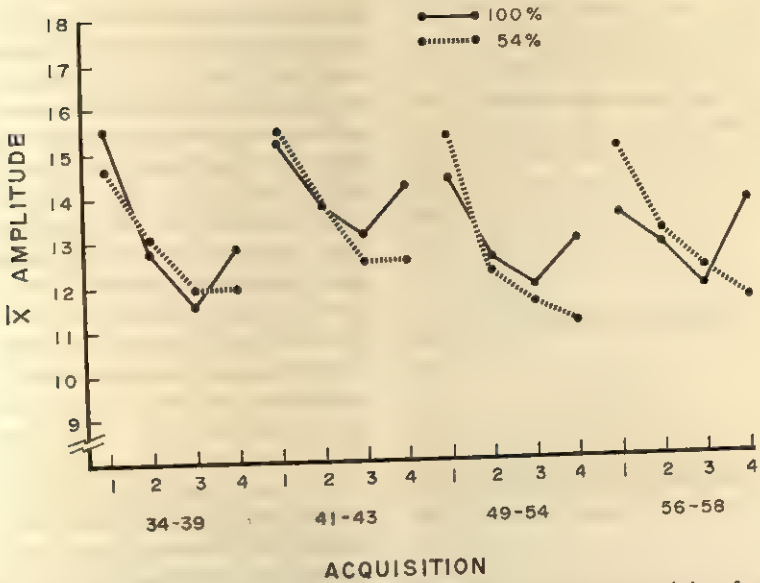

FIG. 5. Mean response amplitude (in mm.) over the last half of training for the two reinforcement schedules (100% and 54%), separately for the four responses within a trial. Each mean is based on the median value of blocks of three trials following reinforced trials.

first three stimuli of each sequence were highly similar for the two conditions, consisting of a decrement from S1 to S3. Presentation of S4, however, resulted in an increment for the 100% condition while the 54% condition resulted in a further decrement. These data were submitted to a four-way analysis of variance with *reinforcement schedule* and *extinction block point* as two between-*S* variables, and *trial blocks* (one to four) and *stimuli* (S3 and S4 only) as two within-*S* variables. Inclusion of extinction block point as a variable, even though this factor was not introduced until extinction, allowed evaluation of differences in training amplitude as a function of subject differences produced by assignment to subsequent extinction conditions. Of 15 resulting F ratios, two were significant at $p < .05$. The *reinforcement schedule* by *stimuli* interaction F was 5.09, $df = 1/54$, confirming the reliability of the differential change in amplitude from S3 to S4 for the two reinforcement conditions. One triple interaction reached significance, the *reinforcement schedule* by *trial* by *extinction block point* F ratio equalling 2.28, $df = 6/162$, $p < .05$. The nature of this interaction, obviously due to sampling fluctuation, is described in Figure 6 which presents amplitude to S3 and S4 separately for each subsequent extinction condition (only the first and last block of training trials are presented in Figure 6). Conditions 1234 and 4 show that 100% *S*s were initially respond-

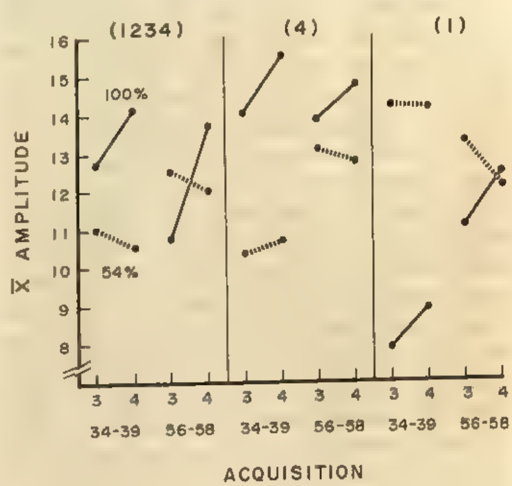

FIG. 6. Mean response amplitude (in millimeters) for trial blocks 34–39 and 56–58 for the two reinforcement schedules (100% and 54%), separately for the last two responses within a trial (3 and 4), and separately for extinction block point (1234, 4, and 1).

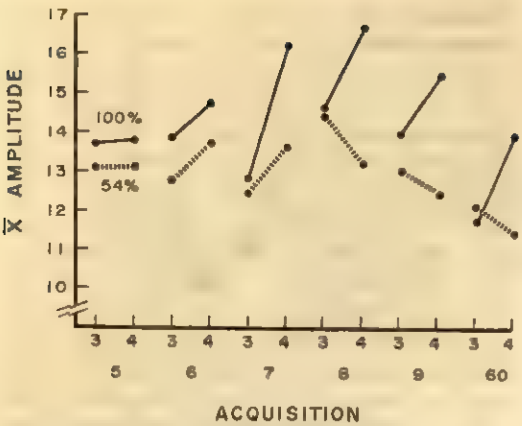

FIG. 7. Mean response amplitude (in millimeters) for training Trials 5–9 and 60 for the two reinforcement schedules (100% and 54%), separately for the last two responses within a trial (3 and 4).

ing with greater amplitude than 54% *Ss* but that this difference decreased, while an opposite trend is indicated for Condition 1. Of greater theoretical importance is the obvious consistency of the increase in amplitude from S3 to S4 for 100% *Ss*: it is apparent on all six of the comparisons in Figure 6, while 54% *Ss* manifest an increase on only one of these comparisons.

If the amplitude goal gradient developed as a result of the 100% reinforcement schedule, one would expect no difference in its form between 100% and 54% conditions early in training, followed by its emergence in Condition 100%. Unfortunately the data of early training trials are contaminated by *E*'s interactions with *S* as previously noted. A careful check of the recordings, however, indicated that beginning with Trial 5, the data of 23 *Ss* from 100% and 24 *Ss* from 54% were free of *E*'s influence; that is, *E* was not in the experimental room. Amplitude of these *Ss* to Stimuli S3 and S4 for Trials 5 to 9 and 60 (the last training trial not preceded by the interuption of program resetting) is presented in Figure 7 separately for the two reinforcement schedules. It is clear that (*a*) there is no difference in the form of the gradient on Trials 5 and 6; (*b*) a differential gradient developed on Trials 7 and 8; and (*c*) it was maintained throughout training (as exemplified by Trial 60). Combining Trials 5 and 6, and 9 and 60, analysis of variance with reinforcement schedule as a between-*S* variable and trials as a within-*S* variable yielded a significant interaction, $F = 4.66$, $df = 1/44$, $p < .05$, indicating that the 100% goal gradient developed over training trials and was absent on early trials.

It was also possible to define amplitude goal gradients on an individual basis and then count the number of *Ss* manifesting such gradients. These analyses yielded one expected and one unexpected result. The expected result was that most of these *Ss* were from the 100% reinforcement condition (16 of 23 *Ss*, $p < .05$). The unexpected finding was that seven of these *Ss* manifested clear gradients which *subsequently disappeared*. It would be tenuous to attribute this to chance: in order to be scored as showing a goal gradient, amplitude to S4 had to be greater than to S3 five trials in succession. Assuming a probability of ½ that such would occur on any one trial, $p = .03$ that such a pattern would occur by chance. Figure 8 presents the mean magnitude of these seven *Ss* to S3 and S4 on Trials 4 to 9, 10 to 30, and 40 to 60. Means on Trials 4 to 9 show slightly smaller amplitudes to S4, thus indicating that the gradients shown on Trials 10 to 30 were developed over trials. The gradient revealed on Trials 10 to 30 is about the same magni-

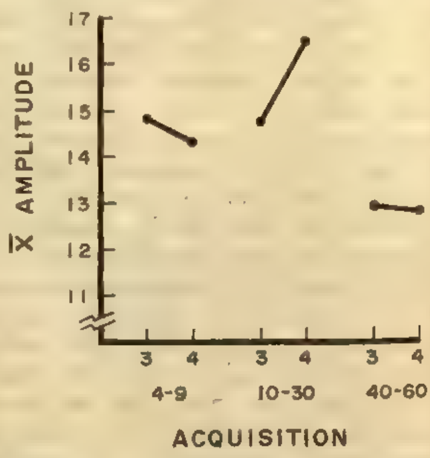

FIG. 8. Mean response amplitude (in millimeters) for seven *Ss* manifesting a "disappearing gradient," for Trials 4–9, 10–30, and 40–60, separately for the last two responses within a trial (3 and 4).

tude as that presented in Figure 7 for 100% *Ss* and is significant as measured by related t, with a value of 4.56, $df = 6$, $p < .01$. During the second half of training, this gradient entirely disappeared for most *Ss*, as indicated by the almost complete lack of a difference between S3 and S4 on Trials 40 to 60. The author is at a loss to explain this phenomenon. Perhaps these *Ss* became so "sure" of the marble that all attendant emotion habituated, leaving a flat gradient. In this respect it may be noted that six of the seven were in the 100% reinforcement condition.

Training Speed

Starting speeds (reciprocals of latency) were computed for the same trials as described previously for amplitude, and are presented in Figure 9. The slow speeds to S1 are due to the time required to place the marble from the previous trial on the board and thus are not comparable to speeds in the remaining segments of the trial. Speeds to S2, S3, and S4 reveal an increase from S2 to S4, with a tendency for speeds to be maximum to S3 late in training. There appears to be little difference in the shape of the curves as a function of reinforcement schedule, but there does appear to be a tendency for 54% *Ss* to respond faster. A four-way analysis of variance, involving the same variables as previously described for the corresponding amplitude data (except that three stimuli, S2, S3, and S4, were included instead of just S3 and S4), yielded only one significant F, that corresponding to the increase in speed from S2 to S4, $F = 4.27$, $df = 2/108$, $p < .025$. It may thus be concluded that starting speeds do not reveal a speed goal gradient which is influenced by reinforcement schedule, but rather that the same gradient is produced by both schedules, consisting of a small but significant rise from S2 to S4, with some tendency for a "peaking" to occur at S3 (a sampling of other trials did not always reveal this peaking).

To determine if the slightly faster speeds of 54% *Ss* were also apparent on early training trials, speeds on Trials 5 to 9 were calculated, using those *Ss* whose data were not contaminated by interactions with E (i.e., the same *Ss* as previously described in the analysis of early amplitude data). These data suggested slower speeds for 54% *Ss*, and hence the late trial speeds for these *Ss* also were determined and are presented in Figure 10 (Trials 60 and 61 have been

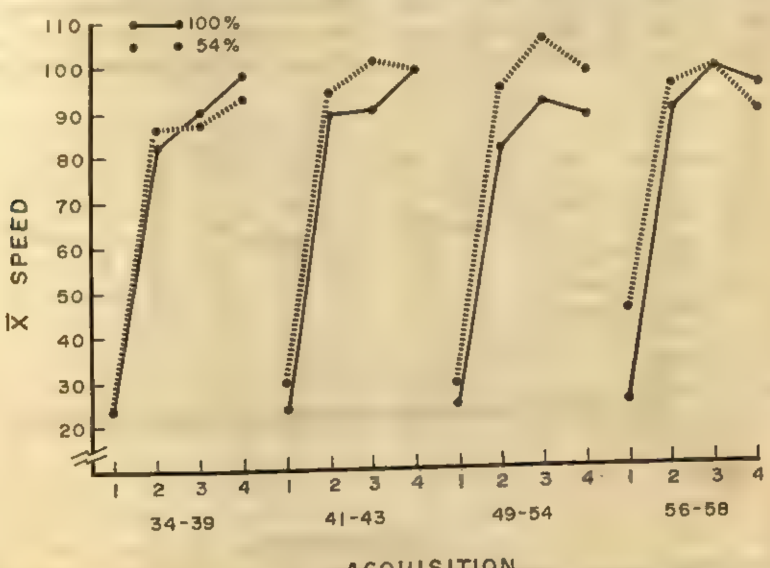

Fig. 9. Mean response speed (100/t sec.) over the last half of training, for the two reinforcement schedules (100% and 54%), separately for the four responses within a trial.

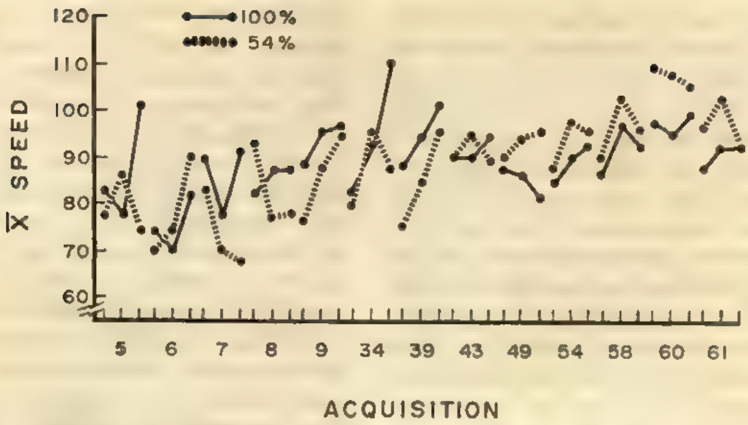

FIG. 10. Mean response speed (100/t sec.) for early (5–9) and late (34–61) training trials, separately for the two reinforcement schedules (100% and 54%) and for the last three responses within a trial (each break represents the end of a trial).

added to illustrate the stability of the trends). The pattern is apparent: on Trials 5 and 6 there is little difference in speeds between the two reinforcement schedules; Trials 7, 8, and 9 reveal faster speeds for the 100% schedule; Trials 34 to 43 suggest a transition period, with no consistent difference between groups; and Trials 49 to 61 reveal faster speeds for the 54% schedule. The first and last five trial blocks in Figure 10 were combined and submitted to analysis of variance, with *reinforcement schedule* as a between-S variable and *trials* and *stimuli* as two within-S variables. Since a number of significant F ratios emerged, a summary table of this analysis is presented in Table 5, and the data are presented in Figure 11, the second and third vertical panels combining trial blocks (Panel 2) as well as reinforcement schedule (Panel 3).

The significant *trials* effect reflects the increase in speed from early to late training trials, while the significant *stimuli* effect reflects a monotonic increase in speed from S2 to S4 (combining both reinforcement schedules, as in Panel 3, there is no evidence of peaking at S3). The significant *stimuli* by *reinforcement schedule* interaction indicates that the form of the speed goal gradient differed for the two reinforcement schedules,

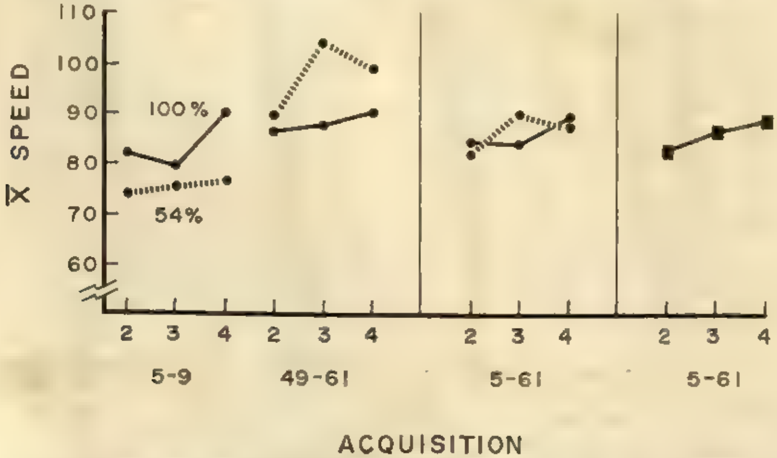

FIG. 11. Mean response speed (100/t sec.) for early (5–9) and late (49–61) training trials, separately for the two reinforcement schedules (100% and 54%) and for the last three responses within a trial (2, 3, and 4). The second vertical panel combines all trials, and the third panel combines reinforcement schedule as well as trials.

TABLE 5

Analysis of Variance of Speeds for Early and Late Training Trials

Source	df	MS	F
Between Ss	95		
Reinforcement schedule (R)	1	2	—
Error	94	292	
Within Ss	480		
Trials (T)	1	1309	13.20**
Stimuli (S)	2	85	4.53*
T × S	2	37	2.45
T × R	1	547	5.52*
S × R	2	59	3.18*
T × S × R	2	21	1.41
Error 1 (T & TR)	94	99	
Error 2 (B & BR)	188	18	
Error 3 (TS & TSR)	188	15	

* $p < .05$.
** $p < .01$.

with a peaking at S3 for 54% and maximum speeds at S4 for 100% (Panel 2). Finally, the significant *trials* by *reinforcement* schedule interaction reflects the differential increase in speeds over training trials for the two reinforcement schedules, with 54% Ss responding slower on Trials 5–9 and faster on Trials 49–61. Related t tests of these within-group changes revealed that the change for 54% Ss was significant ($p < .001$) while the change for 100% Ss was not significant.

Speed data, in summary, show (*a*) a gradual increase over training trials, (*b*) a goal gradient for both reinforcement schedules, (*c*) some hint of a differential gradient, with the 54% schedule producing peaking at S3 and the 100% schedule at S4 (this difference achieving significance when $N = 47$ but not when $N = 60$), and (*d*) 100% Ss responding faster on early trials but 54% Ss responding faster on later trials.

Visual Goal Orientations during Training

Eleven Ss consistently turned their heads away from the stimulus window and towards the marble hose just prior to marble ejection. A notation was made on S's record sheet of this behavior. This visual goal orientation occurred either simultaneous with or just after responding to S4, but before marble ejection a second later. Ten of these 11 Ss were exposed to the 100% reinforcement schedule, yielding a corrected chi-square of 7.12, $df = 1$, $p < .01$. Eight of these Ss were among the 23 manifesting individual amplitude goal gradients as previously defined, resulting in a significant relationship, corrected chi-square of 6.96, $df = 1$, $p < .01$. Thus it appears that Ss who manifested reward anticipations with visual orienting responses also did so with response amplitude and were from the 100% reinforcement condition.

Extinction

Extinction block point (Conditions 1234, 4, and 1) constitutes a second independent variable. Amplitude and speed are discussed in terms of *early extinction* and *late extinction* and are followed by a discussion of *resistance to extinction*. Early extinction refers to those trials during which all Ss responded; that is, it includes all extinction trials up to the point where the first S in that condition stopped responding. Late extinction refers to fifths of extinction and thus samples the entire extinction sequence of all Ss.

Early extinction amplitude. Figures 12, 13, and 14 present mean response amplitude for the three extinction conditions, separately for the two reinforcement schedules. For each condition, extinction trials are extended to that point where the first S stopped responding (12 responses in Condition 1234, 20 in Condition 4, and 20 in Condition 1). In all cases, performance on the last training trial is presented so that changes from training to extinction may be noted. Figure 12 indicates that Condition 1234 resulted in increased amplitude for 100% Ss and in decreased amplitude for 54% Ss. A three-way analysis of variance, with *stimuli* and *trials* as two within-S variables and *reinforcement schedule* as a between-S variable, revealed one significant F, that for the reinforcement schedule by trials interaction, $F = 7.91$, $df = 3/54$, $p < .001$. The increase from Training Trial 61 to Extinction Trial 3 was significant for 100% Ss, $t = 4.4$, $df = 54$, $p < .001$ (utilizing the error term from analysis of variance), while the decrement for the 54% Ss was not significant.

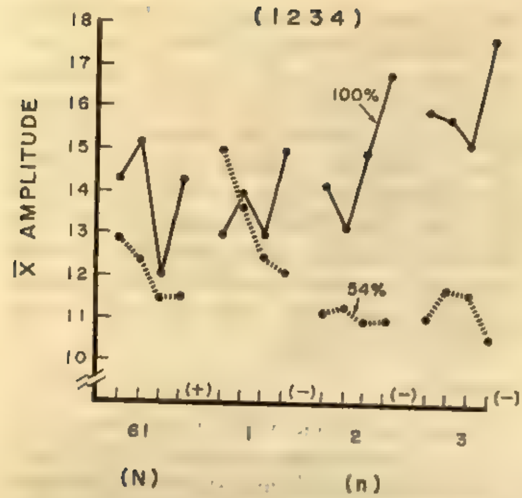

FIG. 12. Mean response amplitude (in millimeters) for the last training trial (61) and the first three extinction trials for Extinction Block Point 1234, separately for the two reinforcement schedules (100% and 54%) and for the four responses within a trial (each break represents the end of a trial).

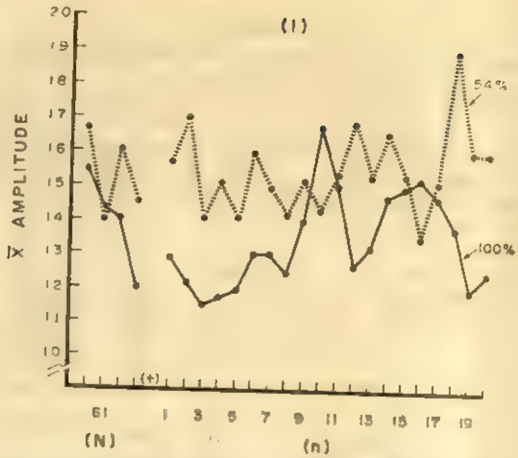

FIG. 14. Mean response amplitude (in millimeters) for the last training trial (61) and for the first 20 extinction responses for Extinction Block Point 1, separately for the two reinforcement schedules (100% and 54%).

Figures 13 and 14 do not reveal similar increments in amplitude for 100% *S*s. In fact, none of the differences between or within groups were significant except for the decrement over extinction trials in Condition 4 for both reinforcement schedules. There is a suggestion that 100% *S*s in Condition 4 manifested an increment in amplitude between Extinction Trials 14 and 20, but statistical evaluation did not verify this observed trend. The early extinction amplitude data could thus be summarized by stating that all groups except 100% *S*s in Condition 1234 showed either no change or a slight decrement in amplitude, while the latter *S*s manifested a temporary increase.

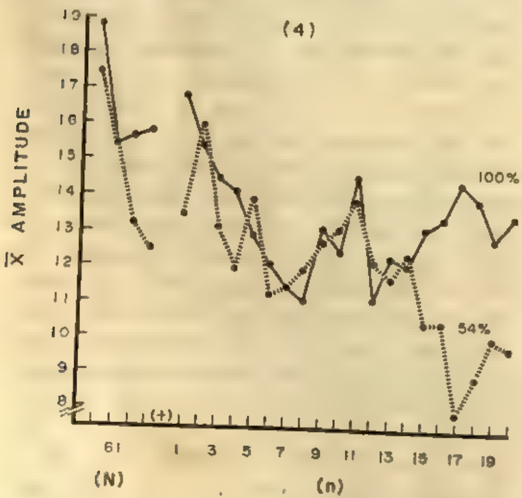

FIG. 13. Mean response amplitude (in millimeters) for the last training trial (61) and for the first 20 extinction responses for Extinction Block Point 4, separately for the two reinforcement schedules (100% and 54%).

Early extinction speed. Figures 15, 16, and 17 present speed data for early extinction trials. Speed to S1 is omitted in Condition 1234, since receipt of the marble on Trials 60 and 61 affected speeds on Trials 61 and the first extinction trial, contributing unwanted variance to the analyses. These figures indicate that speed of 54% *S*s remained fairly constant for all three extinction conditions, but that 100% *S*s in Conditions 1234 and 4 showed a marked drop. Statistical analysis confirmed these trends, indicating that the decrement for the latter two conditions was significant ($p < .05$).

Speed variability was also investigated. Two predictions from frustration theory were tested. First, 100% reinforcement should have produced more approach-avoidance conflict during early extinction than 54% reinforcement, and this difference should have been maximum in Condition 1234, since the similarity of training and

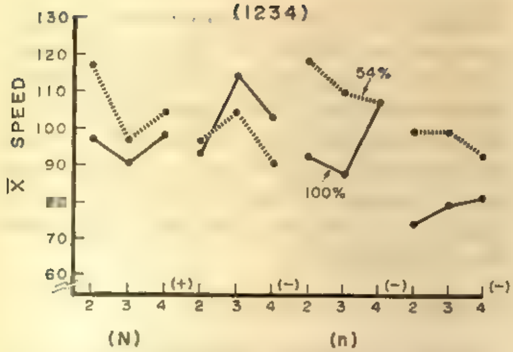

FIG. 15. Mean response speed (100/t sec.) for the last training trial (*N*) and the first three extinction trials for Extinction Block Point 1234, separately for the two reinforcement schedules (100% and 54%) and for the last three responses within a trial (2, 3, and 4).

extinction in this condition would have produced maximum frustration. Second, considering extinction block point, Condition 1234 should have produced maximum conflict and 1 the least, and this difference should have been maximum in the 100% reinforcement conditions, where frustration was presumably maximum.

Both predictions were tested by obtaining variances of changes in speed from training to extinction, thus controlling between-*S*

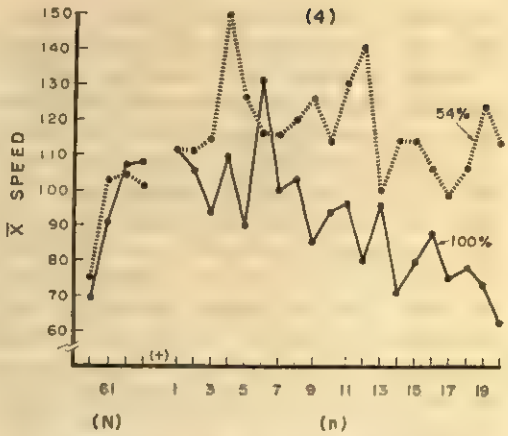

FIG. 16. Mean response speed (100/t sec.) for the last training trial (*N*) and the first 20 extinction responses for Extinction Block Point 4, separately for the two reinforcement schedules (100% and 54%).

differences due to training differences. All *S*s in the 1234 condition, 100% reinforcement schedule (i.e., 1234, 100) were represented for only the first 12 extinction responses (three trials), since the first *S* to extinguish did so at this point. Therefore each *S*'s speed to S2, S3, and S4 on Extinction Trial 3 was summed and subtracted from the corresponding sum on the first

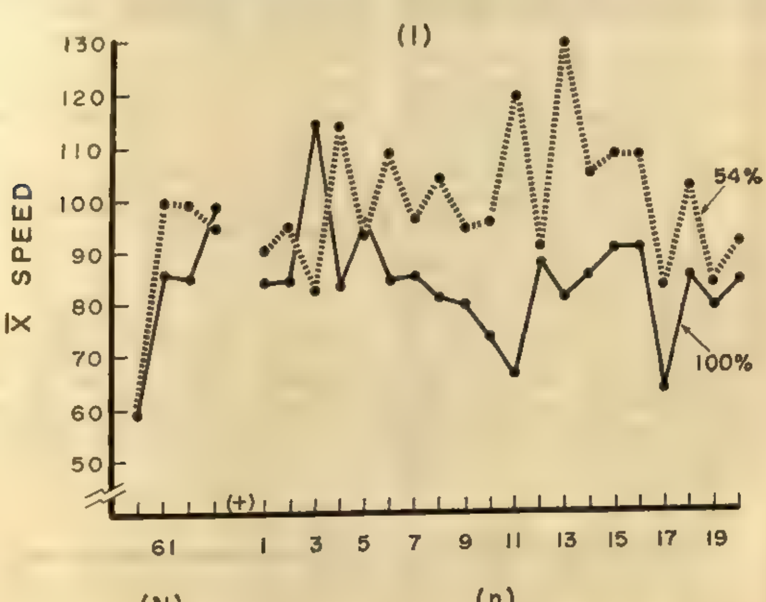

FIG. 17. Mean response speed (100/t sec.) for the last training trial (*N*) and the first 20 extinction responses for Extinction Block Point 1, separately for the two reinforcement schedules (100% and 54%).

extinction trial (S1 speeds were omitted due to depressed speeds on Extinction Trial 1 which resulted from handling of the marble on the last training trial). The variances of these difference scores for Conditions 1234, 100; 1234, 54; and 1, 100 were 23,013, 5,944, and 9,078, respectively. The ratio of the first to the second tests the first prediction, and yields an F of 3.87, $df = 9/9$, $p < .05$. The ratio of the first to the third tests the second prediction, and yields an F of 2.54, $df = 9/9$, $p < .10$. It may thus be concluded that both predictions are supported, the first more clearly than the second.

Late extinction. Figures 18 and 19 present amplitude and speed data, respectively, for the end of training and fifths of extinction. Both figures were evaluated by a three-way analysis of variance, with *reinforcement schedule* and *extinction block point* as two between-*S* variables and *trials* as a within-*S* variable. No significant *F*s were obtained for the amplitude data, although the increase in amplitude for 100% *S*s in Condition 1234 is obvious. Two significant *F*s emerged from the analysis of speed data, with the decrement over trials being significant ($F = 6.30$, $df = 5/270$, $p < .001$), and the interaction of trials and extinction block point approximating significance at $p = .10$ ($F = 1.57$, $df = 10/270$). Although of questionable statistical significance, this latter F is clearly suggested by the speed decrements for Conditions 1234 and 4 and the relative lack of a decrement for Condition 1. These trends are highly similar to speed data for early extinction (see Figures 15–17).

Resistance to Extinction

Table 6 presents mean number of responses to extinction for four different extinction criteria: a response latency of 10, 20, 30, or 60 sec. (the average training latency was about 1 sec.). Six *S*s who responded 750 times and were stopped by *E* were assigned values of 750 if they did not display a latency greater than any one of the four criteria under consideration. For example, if one of these *S*s hesitated 12 sec. on the eighty-ninth extinction response, but did not pause 20 sec. or longer prior to being stopped at $n = 750$, a value of 89 was assigned for the 10-sec. criterion and a value of 750 for the remaining three.

The pattern in Table 6 is remarkably consistent for all four criteria: partial reinforcement produced greater resistance to extinction, and a block point "far" from the

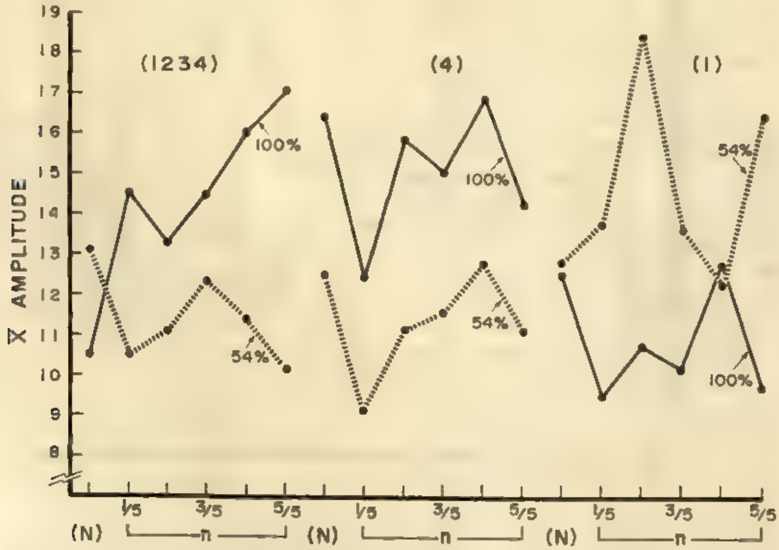

FIG. 18. Mean response amplitude (in millimeters) for the last training trial (N) and fifths of extinction (n), separately for reinforcement schedule (100% and 54%) and extinction block point (1234, 4, and 1).

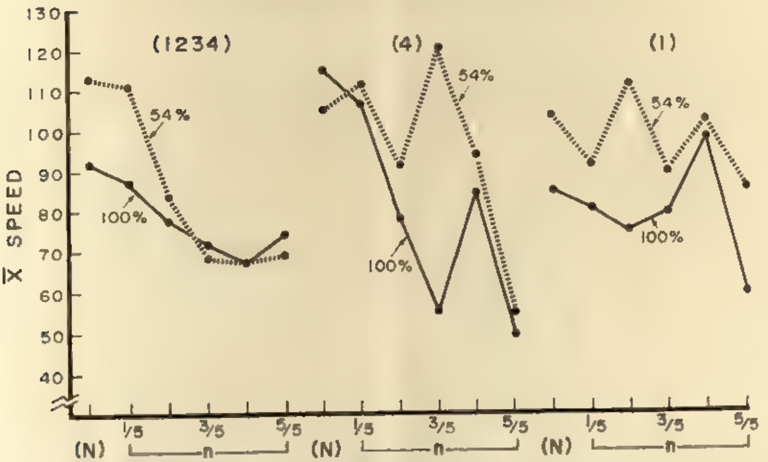

FIG. 19. Mean response speed (100/t sec.) for the last training trial (N) and fifths of extinction (n), separately for reinforcement schedule (100% and 54%) and extinction block point (1234, 4, and 1).

goal (Condition 1) produced greater resistance to extinction than a block point "near" to the goal (Conditions 1234 and 4). Analysis of variance indicated the partial reinforcement effect (PRE) was highly significant ($p < .01$), while extinction block point was of borderline significance, usually yielding an F significant at about $p = .10$, depending upon the criterion measure. As further evidence of its reality, it may be noted that of the six *S*s who responded 750 times, five were in Condition 1. The exact probability of *any* 0, 1, 5, or more extreme split for 6 objects divided into three groups is only .053.

It was previously noted that frequency of amplitude goal gradients and visual goal orientations were both significantly associated with the 100% reinforcement schedule. Since 100% *S*s also extinguished faster, one would expect faster extinction for *S*s manifesting either goal gradients or visual orientation responses. These expectations were confirmed; of the 23 *S*s manifesting amplitude goal gradients, 16 responded fewer than the overall median number of responses to extinction (124 for the 60-sec. criterion), yielding a corrected chi-square of 4.51, $df = 1$, $p < .05$. Nine of the 11 *S*s manifesting visual goal orientations also extinguished with fewer than 124 responses, corrected chi-square of 4.01, $df = 1$, $p < .05$. When schedule of reinforcement was held constant these relationships were not significant, suggesting that the preceding relationships were due to confounding with reinforcement schedule. The question is a moot one, however, since the small *N*s involved in the more refined evaluations resulted in relative insensitive statistical tests. The most that can be safely concluded is that goal gradients and orientation responses were *associated* with rapid extinction.

TABLE 6

MEAN NUMBER OF RESPONSES TO EXTINCTION TO FOUR CRITERIA

Reinforcement schedule	Criterion (sec.)	Extinction block point		
		1234	4	1
100%	10	61	72	201
	20	88	76	220
	30	98	81	236
	60	122	81	237
54%	10	150	162	278
	20	209	230	300
	30	216	246	317
	60	224	250	318

DISCUSSION

Number of responses to extinction indicates that secondary reinforcement effects were not demonstrated; indeed, the opposite was found, with Conditions 1234 and 4 producing faster extinction than Condition 1.

These results are consistent with the findings of Experiments 1 and 2, all three studies showing that presentation of cues previously paired closely with reinforcement (e.g., S+n and S+ in Experiments 1 and 2, S4 in Experiment 3) resulted in faster extinction than presentation of cues more removed from reinforcement (e.g., S− in Experiments 1 and 2, S1 in Experiment 3). As previously noted, such results are consistent with Amsel's frustration theory.

Experiment 3 provides a number of independent tests of Amsel's frustration theory, providing checks on the internal consistency of the data. For example, the training speed data provide clear evidence of slower initial speeds for the 54% reinforcement schedule, but greater final speeds, thus confirming similar findings with animals. According to frustration theory, partially reinforced *S*s are frustrated by the occurrence of nonreinforced training trials, and approach-avoidance conflict results, producing initially slower speeds. With more trials, *S* learns to continue responding in the presence of frustration cues, since such responses are often reinforced, and other responses are not. The avoidance tendencies thus become relatively weak. At the same time, frustration-produced motivation adds to the general drive level of *S*, producing a greater total motivational level than for 100% *S*s. It is the interaction of this higher motivational level with the relatively strong instrumental habit which produces the final greater speed for partially reinforced *S*s. The obtained training speeds are consistent with such an interpretation.

Extinction amplitudes and speeds also provide patterns of relationships which are generally consistent with frustration theory. Condition 1234, 100, for example, would be expected to produce greatest frustration effects early in extinction, since not only were reward expectations stronger for 100% *S*s (recall the amplitude goal gradients), but Extinction Condition 1234 was maximally similar to training conditions. The FE was clearly demonstrated in this condition, consisting of an increment in amplitude and a decrement in speed. Precisely the same results were found in Experiment 2 for Condition S+. Condition 4, 100 would be expected to produce fairly strong frustration effects, too, although the absence of previous stimuli S1–S3 would be expected to work in the opposite direction to some unknown extent. This condition did, in fact, provide most evidence for the FE after 1234, 100, consisting of a nonsignficant increment in amplitude on Extinction Trials 14–20, and a significant decrement in speed during early extinction trials. It is the prediction of these response *patterns* between amplitude, speed, and number of responses to extinction which provides impressive support for frustration theory.

The relationships between another set of response variables are also of interest; specifically, between amplitude goal gradients, visual orienting responses, and resistance to extinction. A basic presupposition of Amsel's theory is that emotional reactions to frustration are a positive function of anticipatory goal responses (rg's) occurring at the time of frustration. To the author's knowledge, all evidence to date offered in support of this assumption is indirect in the sense that rg's have not been directly observed, but have been inferred from stimulus conditions (e.g., Amsel & Hancock, 1957, with animals; Longstreth 1960, 1962, with humans). In the present study the existence of visual orienting responses may be taken as direct instances of such anticipatory goal responses, since that is literally what they were: such orientations were first elicited by presentation of the marble itself, later were elicited by preceding stimuli (S4), and thus became anticipatory.

The isolation of 11 *S*s who consistently made these visual rg's provided test conditions for two theoretical notions. First, Spence has theorized that rg's possess drive properties; they increase the organism's level of motivation (Spence, 1956). Since these rg's occurred shortly before or simultaneous with the joystick response to S4, increased drive might have been expected to manifest itself in the amplitude of this response. In other words, more of these 11 *S*s should have manifested amplitude goal gradients than expected on the basis of the null hypothesis. It will be recalled that eight, in fact, did manifest a gradient, resulting in a significant relationship with

frequency of goal gradients. Spence's theorizing, when extended to human *S*s, was supported.

Second, Amsel's assumption of a positive relationship between rg's and frustration-produced aversive motivation led to the prediction that these 11 *S*s should have extinguished more rapidly than remaining *S*s. It will be recalled that nine of them did so, providing statistically significant support for Amsel's assumption. If, as Spence's theorizing suggests, amplitude goal gradients are a reflection of anticipatory goal responses, a second, independent test of Amsel's assumption was possible: the 23 *S*s exhibiting such gradients could be expected to also extinguish rapidly. Sixteen responded less than the median number of responses to extinction, supporting this prediction ($p < .05$).

Concluding Considerations

Since the reported experiments were interpreted as strongly supporting the application of frustration theory to human behavior, and since the theory was discussed in detail, the final section is mainly concerned with the current status of secondary reinforcement.

There are at least three criteria which must be met when attempting to provide an unambiguous demonstration of Sr effects. One is implicit in the very definition of secondary reinforcement, a second has been discussed in a number of different contexts, and a third is suggested here. After briefly discussing each of these three, it will be argued that *no* so-called demonstration of secondary reinforcement with human *S*s satisfies all three criteria. For purposes of communication, the three precautions are called the *reinforcement* precaution, the *elicitation* precaution, and the *intensification* precaution.

The *reinforcement* precaution is implied in the customary definition of an Sr, to the effect that it must be paired with a reinforcer. The problem is this: a reinforcer may strengthen one response class, but not another—it may not be transsituational. A puff of air, for example, strengthens the conditioned eyeblink, but not necessarily the patellar reflex. A demonstration of secondary reinforcement therefore requires that evidence be presented showing that the original reinforcer does in fact strengthen the response which is to be followed by only Sr in the test phase. Failure to provide such evidence creates a double-edged sword: if Sr effects are not demonstrated, it can be argued that no evidence was presented demonstrating that Sr was paired with a reinforcer *appropriate for that response* in the first place, and hence the test was inadequate; if Sr-like effects are demonstrated, it can be argued that pairing Sr with a reinforcer was not necessarily the critical variable, since no such reinforcer was demonstrated, and hence other interpretations become reasonable.

The *elicitation* precaution, elsewhere identified with Bitterman's discrimination hypothesis (Myers, 1958), involves one such interpretation. According to this requirement, it must be insured that the Sr does not *elicit* the response it is assumed to *strengthen*. It is for this reason that Skinner box studies of Sr effects have been considered to be of questionable value: the click of the food magazine, previously paired with food pellets and then presented alone following a bar press, may elicit the next bar press rather than strengthen the last one (e.g., Bulgeski, 1956). A highly related problem is that the click may elicit approach responses to the food cup, which then keep *S* in the vicinity of the response bar, thus increasing the probability that a press will occur (e.g., Wyckoff et al., 1958). In this case it may be said to *indirectly* elicit the response it presumably strengthens. In either case, secondary reinforcement processes are clearly not required: a simple cueing effect accounts for the data.

The *intensification* problem is suggested by discovery of the frustration effect: that removal of a reinforcer from its customary spatio-temporal location strengthens immediately subsequent behavior. The crucial consideration is that precisely the same behavior can be strengthened by either subsequent presentation or preceding absence of reinforcement. Thus Amsel and others have shown in a number of double-alley studies that alley running initially learned by reinforcement in the goal boxes is further

strengthened in Alley Two by the omission of reinforcement in Goal Box One. Experiments 2 and 3 in the present monograph have demonstrated a similar phenomenon with human *Ss*, as have other studies (Haner & Brown, 1955; Holton, 1961; Longstreth, 1965; Ryan, 1965). As most Sr studies involve the removal of a reinforcer, but presentation of most, if not all, other cues previously paired with reinforcement, conditions for the FE are maximal. Strengthening of such behavior is obviously due to *preceding* cues, and to that extent cannot possibly be due to Sr effects, which are exerted *after* the response.

Turning, then, to Sr studies with human *Ss*, it may first be noted that a shift in the ratio of positive to negative reports occurred around 1960. Reviewing the literature to 1958, Myers concluded, "The author feels that secondary reinforcement is inadequately defined and inadequately demonstrated.... [Myers, 1958, p. 299]." Since then, Myers and his associates have reported a series of studies which seem to tip the scales in the positive direction: the author is aware of nine published studies by this group, seven of which report Sr effects (Fort, 1961, 1965; Leiman, Myers, & Myers, 1961; Myers, 1960; Myers, Craig, & Myers, 1961; Myers & Myers, 1962, 1963, 1964, 1965). It may be noted in passing, however, that other recent studies are not so confirmatory. Ignoring the negative results in the present three studies, two others report negative results (Kass, Wilson, & Sidowski, 1964; Longstreth, 1962), while one reports positive results (Sidowski, Kass, & Wilson, 1965). Further, the author is aware of three unpublished studies, all of which report negative results (Donaldson, 1961; Estes, 1960; Hall, 1964).

It would seem, then, that the Myers studies afford the best opportunity to learn what unique procedures produce consistent Sr effects. Examination of the first one in the series will prove as instructive as any, since it is concluded, "The results clearly indicate that tokens can be established as strong secondary reinforcers for preschool children [Myers, 1960, p. 177]." Preschool children were exposed to the following sequence: light—button press—receipt of token—insertion of token—button press—candy. After 20 "conditioning" trials of this sequence, *Ss* were extinguished under one of the two following sequences: (*a*) light—button press—button press—etc. (i.e., neither token nor candy was delivered), (*b*) light—button press—receipt of token—insertion of token—button press—etc. (i.e., the token was delivered, but not candy). It was found that Condition (*b*) resulted in significantly more responses during the 3⅓ min. of "extinction," leading to the quoted conclusion.

Now, what happened in this situation? To begin with, no acquisition data are presented to document the claim that "conditioning" occurred. Thus, there is no evidence that candy was a reinforcer for the button-press response. Indeed, a control group did *not* receive candy, and yet apparently performed the same as other groups, since it finished the 20 training trials in the same length of time. The study actually presents evidence, then, that the candy played no role in acquisition of the response which was subsequently used to test for Sr effects, and therefore, by definition, the candy was not a reinforcer. Thus the reinforcement precaution was clearly not satisfied in this study.

Next, it may be noted that during training, receipt of the token and its insertion regularly preceded button pressing, and thus presumably became functional as cues eliciting the button-press response. Presentation of these cues during extinction, as in Condition (*b*), would therefore be expected to result in more button-press responses merely on the basis of this previous S-R association. Clearly, then, the tokens may have elicited the next response rather than strengthened the preceding response, particularly in view of the previous point that the candy was apparently not a reinforcer to begin with. The authors seem to have become aware of this possibility a year later, since in a very similar study it is stated, "The Sr, when administered in extinction, seemed to release the additional behavior in the chain... [Myers et al., 1961, p. 771]." If it "released" the next response, there is no need to assume it

strengthened the preceding response, and thus it may be concluded that the elicitation precaution was not satisfied.

Finally, it may be noted that failure to obtain candy in Condition (*b*) was presumably more frustrating than failure to obtain candy in Condition (*a*), since (*b*) contained more of the cues previously associated with candy and hence presumably elicited stronger anticipations of candy. Since extinction involved a free operant procedure rather than a discrete trial procedure, *S* was able to respond shortly after frustration occurred, thereby maximizing the probability that perseverative frustration effects were operative during subsequent responses. If one of these effects was an energizing one, as found in the present studies, then the higher rate of responding under Condition (*b*) may have been a frustration effect rather than an Sr effect. It is unfortunate that extinction did not extend beyond 3⅓ min., since the high response rate may have been predictive of rapid extinction, just as the present studies found high response *amplitude* to be associated with rapid extinction. It must be concluded, then, that the intensification precaution was also not satisfied in this study.

An examination of every available published study reporting Sr effects with human *S*s reveals that at least one of these precautions was not met, although it is unusual to find a study which violated all three. It is on this basis that it becomes reasonable to argue that the phenomenon has not yet been clearly demonstrated with human *S*s. Space does not permit a similar evaluation of the infrahuman literature. A careful look at these studies may well lead to a similar conclusion. Such a possibility is undoubtedly what Amsel had in mind when he wrote,

> It occurred to me ... that there is a similarity of operations ... raising the possibility that at least some of the effects which we have been attributing to secondary reinforcement may actually depend upon the arousal of frustration and its reduction [Amsel, 1961, p. 35].

SUMMARY

The conditions which allegedly result in the development of a secondary reinforcer (Sr) are highly similar to those which result in frustration: pair a neutral stimulus with reinforcement a number of times and then present it alone. According to the notion of secondary reinforcement, such a cue will acquire the function of reinforcement. Its presentation following a response should therefore strengthen that response in the same manner that a "primary" reinforcer does. Such strengthening is called the secondary reinforcement effect (Sr effect).

According to Amsel's frustration theory, such a cue is not reinforcing, but on the contrary, elicits an unconditioned aversive emotional response R_F. The effect of R_F upon the modification of preceding or subsequent behavior is described by Amsel's theory. In certain situations such effects are easily distinguished from Sr effects; in other situations the distinction is more difficult, leading to confusions about the roles of secondary reinforcers and R_F. Indeed, it is even possible that one of the two concepts is sufficient to account for the data often ascribed to the other. This monograph reports a series of studies designed to investigate the roles of these concepts in human instrumental conditioning.

Experiment 1 paired two stimuli (S+ and n) with reinforcement and another stimulus (S−) with nonreinforcement. One-third the *S*s were subsequently presented with both Sr's but no reinforcement following a joystick response (Condition S+*n*), another third with one of these stimuli (Condition S+), and another third with S−. Condition S+*n* resulted in fastest extinction, the greatest decrement in speed following the very first extinction trial, and greatest speed variability early in extinction. These results did not support secondary reinforcement theory, but were consistent with frustration theory except that (*a*) differences between Conditions S+ and S− were generally not significant and (*b*) response amplitude changes predicted by frustration theory were not observed. Experiment 2 replicated conditions S+ and S− and introduced some procedural changes designed to further investigate the relative applicability of frustration versus secondary reinforcement concepts. The results were completely

in accord with frustration theory, the S+ condition resulting in faster extinction, a temporary decrement in speed, greater variability in speed, and a temporary increment in response amplitude.

Experiment 3 approached the problem from a different viewpoint, introducing two different variables, schedule of reinforcement (100% and 54%) and "nearness to goal" at the time of extinction. A number of relationships emerged from this study, which may be summarized as follows: (*a*) the 100% reinforcement schedule resulted in the appearance of amplitude and speed goal gradients, defined as increases in amplitude and speed as the goal was approached; (*b*) the 100% condition also produced more anticipatory orienting responses toward the goal locus than the 54% condition; (*c*) the 54% condition resulted in slower speeds early in training but in faster final speeds; (*d*) the PRE was observed; (*e*) faster extinction was obtained for *S*s extinguished "near" to the goal than for *S*s extinguished "far" from the goal; (*f*) conditions producing the fastest extinction also produced temporary decrements in speed, increments in amplitude, and greater variability in speed; and (*g*) *S*s manifesting goal gradients and anticipatory orienting responses during training extinguished faster than other *S*s. These results were remarkably consistent with predictions from frustration theory. Sr effects were again not demonstrated.

Final considerations were concerned with an evaluation of evidence from other experiments which had been interpreted as suggesting Sr effects. Three precautions were discussed which seemed important in the evaluation of these studies. Viewed in this light, it was concluded that Sr effects have not yet been clearly demonstrated with human *S*s.

REFERENCES

AMSEL, A. The role of frustrative nonreward in noncontinuous reward situations. *Psychological Bulletin*, 1958, **55**, 102–119.

AMSEL, A. Hope comes to learning theory. *Contemporary Psychology*, 1961, **6**, 33–36.

AMSEL, A. Frustrative nonreward in partial reinforcement and discrimination learning: Some recent history and a theoretical extension. *Psychological Review*, 1962, **69**, 306–328.

AMSEL, A. Partial reinforcement acquisition and extinction effects under within-subject and between-subject conditions. Paper read at Psychonomic Society, Niagara Falls, 1964.

AMSEL, A., & HANCOCK, W. Motivational properties of frustration: III. Relation of frustration effect to antedating goal factors. *Journal of Experimental Psychology*, 1957, **53**, 126–131.

BITTERMAN, M. E., FEDDERSEN, W. E., & TYLER, D. W. Secondary reinforcement and the discrimination hypothesis. *American Journal of Psychology*, 1953, **66**, 456–464.

BROWN, J. S. *The motivation of behavior.* New York: McGraw-Hill, 1961.

BUGELSKI, B. R. *The psychology of learning.* New York: Holt, 1956.

CASTANEDA, A., & WORELL, L. Differential relation of latency and response vigor to stimulus similarity in brightness discrimination. *Journal of Experimental Psychology*, 1961, **61**, 309–314.

DONALDSON, T. E. Secondary reinforcement vs other stimulus effects in extinction and learning of a new response. Unpublished doctoral dissertation, Purdue University, 1961.

EGGER, M. D., & MILLER, N. E. Secondary reinforcement in rats as a function of information value and reliability of the stimulus. *Journal of Experimental Psychology*, 1962, **64**, 97–104.

ELAM, C. B., TYLER, T. W., & BITTERMAN, M. E. A further study of secondary reinforcement and the discrimination hypothesis. *Journal of Comparative and Physiological Psychology*, 1954, **47**, 381–384.

ESTES, R. K. Number of training trials and variations in stimulus cues as factors effecting extinction in children. Unpublished master's thesis, University of Southern California, 1960.

FINGER, F. W. Quantitative studies of "conflict": I. Variations in latency and strength of the rat's response in a discrimination-jumping situation. *Journal of Comparative Psychology*, 1941, **31**, 97–127.

FORT, J. G. Secondary reinforcement with preschool children. *Child Development*, 1961, **32**, 755–764.

FORT, J. G. Discrimination based on secondary reinforcement. *Child Development*, 1965, **36**, 481–490.

HALL, J. Secondary reinforcement or frustration? Unpublished doctoral dissertation, University of Southern California, 1964.

HANER, C. F., & BROWN, P. A. Clarification of the instigation to action concept in the frustration-aggression hypothesis. *Journal of Abnormal and Social Psychology*, 1955, **51**, 204–206.

HOLTON, R. B. Amplitude of an instrumental re-

sponse following the cessation of reward. *Child Development,* 1961, **32,** 107–116.

Kass, N., Wilson, H., & Sidowski, J. B. Effects of number of training trials upon the development of a secondary reinforcer with children. *American Psychologist,* 1964, **19,** 451 (Abstract).

Lambert, W. W., Lambert, E. G., & Watson, P. D. Acquisition and extinction of an instrumental response sequence in the token-reward situation. *Journal of Experimental Psychology,* 1953, **45,** 321–326.

Leiman, A. J., Myers, J. L., & Myers, N. A. Secondary reinforcement in a discrimination problem with children. *Child Development,* 1961, **32,** 349–353.

Longstreth, L. E. The relationship between expectations and frustration in children. *Child Development,* 1960, **31,** 667–671.

Longstreth, L. E. Incentive stimuli as determinants of instrumental response strength in children. *Journal of Comparative and Physiological Psychology,* 1962, **55,** 398–401.

Longstreth, L. E. An operational distinction between secondary reinforcement and frustration. *American Psychologist,* 1964, **19,** 452 (Abstract).

Longstreth, L. E. Unconditioned and conditioned frustration in retardates. *American Psychological Association Proceedings,* 1965, 1–2.

McKeever, B., & Forrin, B. Secondary reinforcing properties of informative and non-informative stimuli. *Psychonomic Science,* 1966, **4,** 115–116.

Mowrer, O. H. *Learning theory and behavior.* New York: Wiley, 1960.

Myers, J. L. Secondary reinforcement: A review of recent experimentation. *Psychological Bulletin,* 1958, **55,** 284–301.

Myers, J. L., & Myers, N. A. Secondary reinforcement in children as a function of conditioning associations, extinction percentages. *Journal of Experimental Psychology,* 1964, **68,** 611–612.

Myers, J. L., & Myers, N. A. Secondary reinforcement in children as a function of conditioning associations, extinction percentages and stimulus types. *Journal of Experimental Psychology,* 1963, **65,** 455–459.

Myers, N. A. Extinction following partial and continuous primary and secondary reinforcement. *Journal of Experimental Psychology,* 1960, **60,** 172–179.

Myers, N. A., Craig, G. J., & Myers, J. L. Secondary reinforcement as a function of the number of reinforced trials. *Child Development,* 1961, **32,** 765–772.

Myers, N. A., & Myers, J. L. Effects of secondary reinforcement schedules in extinction on children's responding. *Journal of Experimental Psychology,* 1962, **64,** 586–588.

Myers, N. A., & Myers, J. L. A test of a discrimination hypothesis of secondary reinforcement. *Journal of Experimental Psychology,* 1965, **70,** 98–101.

Ratner, S. C. Reinforcing and discriminative properties of the click in a Skinner box. *Psychological Reports,* 1956, **2,** 332.

Ryan, T. J. The effects of nonreinforcement and incentive value on response speed. *Child Development,* 1965, **36,** 1067–1082.

Schoenfeld, W. N., Antonitis, J. J., & Bersh, P. J. A preliminary study of training conditions necessary for secondary reinforcement. *Journal of Experimental Psychology,* 1950, **40,** 40–45.

Sidowski, J. B., Kass, N., & Wilson, H. Cue and secondary reinforcement effects with children. *Journal of Experimental Psychology,* 1965, **69,** 340–342.

Spence, K. W. *Behavior theory and conditioning.* New Haven: Yale University Press, 1956.

Wagner, A. R. Effects of amount and percentage of reinforcement and number of acquisition trials on conditioning and extinction. *Journal of Experimental Psychology,* 1961, **62,** 234–242.

Wyckoff, L. B., Sidowski, J., & Chambliss, D. An experimental study of the relationship between secondary reinforcing and cue effects of a stimulus. *Journal of Comparative and Physiological Psychology,* 1958, **51,** 103–109.

(Received August 26, 1965)

Vol. 80, No. 12 Whole No. 620, 1966

Psychological Monographs: General and Applied

SHORT-TERM MEMORY IN THE MENTALLY RETARDED: AN APPLICATION OF THE DICHOTIC LISTENING TECHNIQUE[1]

ALDRED H. NEUFELDT

Psychiatric Services Branch, Department of Public Health, Regina

A series of experiments was conducted to investigate short-term memory (STM) in mental retardates, utilizing the dichotic listening technique initiated by Broadbent (1958). The primary purpose of these experiments was to discern whether or not STM capacity and/or strategy of encoding information could account for some of the differences between retardates and normals. Four groups of Ss were compared: two groups of retardates, one organic (O) and one cultural familal in nature (F), a normal mental age control group (NMA), and a chronological age control group (NCA). In sum, the evidence indicated that STM capacity was indeed an important difference between retardates and group NCA, though the capacity of Groups O, F, and NMA was essentially the same. Furthermore, both normal groups demonstrated a marked degree of flexibility in their adaptation of different strategies of recall to various rates of informational input, and ability in using more ambiguous strategies. Such flexibility was not found in the retardates. Differences between the two normal control groups, on the other hand, were indicative of the degree to which both memoric capacity and ability to make use of useful strategies develops in normal individuals over time. The implications of these results were discussed.

One of the best ways to learn about a system of which little is known is to study that system at its points of breakdown. Applying this principle to the topic at hand it soon becomes apparent that one of the most promising loci of investigating short-term memory (STM)[2] would be the mentally retarded. The question as to why the mentally retarded are retarded can be looked at from two points of view—either theirs is a problem of information *retrieval*, or one of information *acquisition*. If one holds that it is one of retrieval, this suggests that the retarded can encode[3] information as well as normals but are not able to evoke that information again—essentially an untestable hypothesis. The second proposal, that the problem is one of acquisition, suggests that the retarded are *not* able to encode as much as normals, or at least are not able to retain such information long enough for it to be permanently stored —a problem of STM and hence potentially testable. The experiments presented, then, deal with the mentally retarded and have been designed to elucidate some of the concepts of immediate memory.

Memory and learning. Though it has generally been conceded that memory per se forms an integral part of any learning situation, psychologists on the whole have traditionally espoused relatively little interest in the former as compared to the latter. This unilateral emphasis has slacked off considerably since the mid-50's, however. The cogency with which G. A. Miller (1956) and

[1] The research presented here was carried out in partial fulfillment of the PhD degree at the University of Hawaii, Honolulu. The author expresses his appreciation in particular to Ronald C. Johnson for his mental stimulation and evaluative criticism.

[2] Short-term memory can roughly be distinguished from long-term memory as that memory lasting but a few seconds or minutes as compared to days and weeks. A common example of short-term memory in action is the retention of a telephone number. We remember it from the time we look it up until it has been dialed, but seldom longer provided, of course, we do not have to focus our attention on something else in between.

[3] The term "encoding" is used here to refer to the taking in of information by the organism. Osgood (1957) would refer to such a process as "decoding." In view of the literary definition of the prefix "en," however, it was felt that the former had the more proper connotation. This usage of "encoding" agrees closely with that of James Deese (1958, p. 247).

D. E. Broadbent (1958) have presented the case for STM now has even the traditionally cautious functionalist considering this facet of the study of memory and learning. The importance which this work has assumed has been noted by Melton (1963), and something of its nature can be seen in reviews by Posner (1963) and Postman (1964).

The rise of psychological information theory (cf. Berlyne, 1957) presented a major development with which to explore behavior. Viewing the organism in terms of information (stimulus) flow led to consideration of the organism in terms of memoric capacity. In this fashion STM has come to be taken as of major importance in human information-processing. Information, whatever the source, upon entering the organism (in the case of exteroceptor stimulation) presumably enters an STM system. Such information may either be lost here due to spontaneous decay or to interference from other incoming stimuli, or both (which, or both is a theoretical issue still very much alive), or it may be transferred to some permanent storage locus (long-term memory). The import of STM to any consideration of learning thus immediately becomes apparent in that an STM system can control what and how much information the organism encodes.

Short-term memory. Most theorists, such as Osgood (1953, 1957), Broadbent (1958), and Miller, Galanter, and Pribram (1960), who consider learning and performance in terms of STM would picture a learning situation something as follows: the organism is faced with a task which is to be learned. During the learning process, the organism is bombarded with large amounts of information in a short period of time. Now, presumably, the organism which is able to store the *most relevant* information long enough for it to be transferred to the long-term memory system learns most. The learning problem thus becomes one of immediate storage capacity. Two factors which might affect an organism's effective storage capacity become apparent: (*a*) organisms may differ in inherent short-term *storage capacity* and (*b*) organisms may differ in *strategy* of encoding the available information, some strategies being more optimal than others (cf. Bruner, 1957; Neufeldt, 1963).

Particularly germane to the question of capacity and strategy is the problem of individual and group differences as these should be readily amenable to interpretation in terms of STM. Consider differences between fast and slow learners, between normals and mentally retarded, or merely the effect of increasing age on learning performance in normal individuals. These are all problems potentially interpretable in terms of STM, yet traditional measures, such as the digit span, have revealed few differences between such overtly distinguishable groups. The immediate storage capacity of these various groups appears to be very much the same, a point rather well made by Miller (1956). The major point espoused by Miller is not, though, that the memory spans of different individuals differ little but that better or poorer use can be made of the span one has. A relatively efficient strategy of encoding or recoding information, for instance, can increase the apparent memory span, storing more information than a relatively inefficient or poor strategy can. Considerable evidence in support of the importance of encoding strategies is available (cf. Neufeldt, 1963).

Besides the importance of strategies on increasing the apparent capacity of STM, however, it is also possible that inherent differences of capacity do exist, though not clearly delimited by the traditional tests of memory span. Consider the effect of increasing age on learning performance in normal people as a case in point. Results of various experiments would appear to be in conflict, some showing very little falling off with increasing age, others a considerable amount (cf. Welford, 1958, pp. 247 ff.). Part of such disagreement is in all probability due to a lack of consistency in methods of administration from test to test, or from time to time on a given test for that matter. In addition, however, the techniques involved often are obtaining only a gross estimate of STM capacity, thus allowing for greater variance of results. A tool which appears to overcome some of these difficulties has been presented by Broadbent (1954). This modified memory-span technique, termed "dichotic listening," has proven of considerable potential utility in the analysis of such ca-

pacity. For example, after equating elderly patients with and without gross memory disorder on digit span, Inglis and Sanderson (1961), using this technique, nevertheless found that differences in capacity still did exist, as is to be expected.

It would thus appear that some of the traditional measuring techniques (such as digit span) may, for the reasons specified above, not be either as useful or as accurate as the dichotic-listening technique described. Such would seem to be the case in the study of STM in the aged at least (cf. Caird & Inglis, 1961; Inglis, 1957; Inglis & Sanderson, 1961). The problem of whether or not these results are generalizable to other groups who differ in learning ability, such as fast versus slow learners, or normals versus mentally retarded, remains to be tested. As a measure of short-term storage capacity, the dichotic listening technique would seem to be a highly sensitive method of getting at such differences. Two questions of some importance arise: First, can this dichotic task also be used as an index of encoding strategy? If it can tap strategy as well as capacity, then we have, indeed, a useful device. Second, can this technique tell us anything about the structure of STM? To answer questions of this nature, we should consider the theorizing of Broadbent in some detail.

The Broadbent model. Broadbent's *Perception and Communication* (1958) has played a key role in the rapid development of interest in STM. Much of the experimental data marshaled by Broadbent in support of his approach has been derived from the dichotic listening technique already mentioned. In a typical dichotic listening experiment a subject (*S*) listens to two sequences of digits presented in such a way that one number arrives at the left ear at the same time that a different number arrives at the right; for example, the left hears 637 while the right hears 194 in such a fashion that as the left hears "6," the right hears "1," etc. Broadbent (1954) discovered that if such pairs of digits are presented in rapid succession—that is, at the rate of two pairs a second—*S*, when required to identify the material heard in a free recall manner, will tend to report first all of the digits presented to one ear and then the digits presented to the other (either 637194 or 194637 for the above example). He also found that when *S* is required to report the first pair of digits (e.g., 61) first, then the second pair, and then the third (in the case where three pairs of digits have been presented), recall is less successful than when *S* is permitted to give the digits heard in one ear and then those heard in the other. The first finding has been confirmed by Bryden (1962); the second, with slight modification, by Moray (1960). It is thus evidently easier to report the digits ear by ear than to report them pair by pair so long as rate of presentation is fairly rapid. On the other hand, when the material is presented slowly, it is more common for *S*s to give the material in the order of arrival (Broadbent, 1954; Bryden, 1962).

Broadbent (1958) has proposed a model in terms of sensory "channels" to account for these findings. He argues that the material arriving at one ear (channel) is attended to and perceived as it arrives, while the material arriving at the other ear is held in short-term storage. Once *S* has perceived all the material on the first channel, he can attend to the material from the second, provided that the memory traces in the short-term storage system have not decayed. The *S* is unable to switch attention from ear to ear (channel to channel) fast enough to assimilate all the incoming information when a rapid rate of presentation is used and so "listens" (attends) to one ear while a "filter mechanism" of some sort shunts the information coming into the other channel into storage; thus, *S* reports all the numbers from the one ear first. If the rate of presentation is slowed down, *S* has enough time to shift attention channel to channel, and so can report the material in the order of arrival. Bryden (1962, 1964) has considered these various modes of recall in terms of strategy of recall.

The exact nature of the "filter mechanism" is left unspecified by Broadbent, and a consideration of its nature and locus has resulted in some disagreements (cf. Emmerich, Goldenbaum, Hayden, Hoffman, & Treffts, 1965; Moray, 1960). There is, however, fairly good agreement that the STM system must contain at least two parts: (*a*) a

limited capacity channel which attends to information as soon as it is received (termed perceptual, or P system); and, (*b*) a storage area which can store, for short periods of time, such superfluous information as cannot immediately be carried by the P system (termed S system). The dichotic listening studies of Broadbent (1954, 1956, 1957), of Broadbent and Gregory (1961), Moray (1960), and Bryden (1962, 1964) support these concepts as outlined.

The dichotic technique thus seems to be an excellent indicator of memory capacity, both of the S and P systems. Broadbent's discussion of order of recall, however, would have it that the structure of STM is what determines whether a person recalls the digits by ear or in temporal order. That is to say, the switching mechanism is what determines *S*'s order of recall, and as was noted earlier, considerable disagreement with Broadbent has arisen over this claim. In view of the evidence, a better approach, it seems, would be to accept STM as a two-part system, but to view, along with Yntema and Trask (1963), *S*'s recall performance more in terms of a search process, and, along with Bryden (1962), the order of recall as a strategy. The fact that even at fast rates of presentation intrusions from the second ear do occur in recall of the first, and vice versa (what Bryden would term "attempted ear order") would indicate that the two channels are not totally separable as suggested by Broadbent. However, the ear order of recall may well be the best *strategy* of recall that *S* can adopt.

The problem of the retarded. From the studies of Broadbent (1954, 1956, 1957) Inglis (1960), Inglis and Sanderson (1961), and Bryden (1962, 1964) the utility of the dichotic listening technique in studying STM would seem widespread. It is capable of picking up differences where conventional techniques fail, and if differences are present, as in the case of senile versus intact subjects, can pick up these differences with a very small population sample. The results appear to be generalizable to channels of input other than ear alone (cf. Broadbent, 1957; Caird & Inglis, 1961). Research using this technique has led to renewed interest in attention and also to the postulation of a theoretical structure of STM (Broadbent, 1958) that has generated a considerable amount of research. Finally, as a technique, it is amenable not only to the discovery of differences in memory capacity, but also to the identification of strategies used in encoding information (Bryden, 1962, 1964), a problem most other techniques used in the investigation of STM find difficult to handle.

Whether or not differences in storage capacity and/or strategy of information coding will account for group differences, such as those between normals and retardates, remains to be seen. The hypothesis of this paper is, however, that these two concepts should go a long way towards the explanation of such differences as do exist. As was indicated at the outset of this chapter, the problems of mental retardation should make a good proving-ground for such an hypothesis.

N. R. Ellis (1963) has pointed out that really very little is known about the STM of mentally retarded—either in terms of strategy or capacity. On a gross level distinctions can be made between the organic and cultural-familial retardates for instance (cf. Robinson & Robinson, 1965). When, however, attempts have been made to test for such differences with such short-term techniques as delayed recall, relatively few have been found (cf. Osborn, 1960; Weatherwax & Benoit, 1957). It has furthermore been observed that retardates as a whole do about as well on laboratory tasks under some circumstances as do normal *S*s. For example, Shapiro and Johnson[4] have found that mentally retarded do quite well on laboratory learning tasks so long as learning trials are distributed and extraneous "noise" in the learning system is minimal. Laboratory tasks of this nature depend, of course, heavily on STM. In a task such as mentioned above, the amount of information to be handled by the *S*s is limited, it can be handled successively, and most of it is relevant. In a scholastic setting, however, the information available to the individual is almost infinite, and only some

[4] G. M. Shapiro and R. C. Johnson. The effect of massed vs. distributed practice on the learning of bright, average and dull children matched in mental age. Unpublished manuscript, Honolulu, 1964.

of it is relevant to the task of learning. Differences in learning in such a case could be thought of in terms of coding strategy—retardates use less optimal strategies than do normals, such as attempting to encode all the information available, or not discriminating between relevant and irrelevant information, and so forth. If this is the case, such differences should become evident with the dichotic listening task where two different sources of information are present, both of which are highly demanding of attention and where the information is presented too rapidly to be handled successively. In such a situation normals (at least those above chronological age [CA] 11, as demonstrated by Inglis & Caird, 1963) tend to report all the information received by one ear before reporting that of the other. How retardates respond in such a situation was not known and remained for the following experiments to discern. One might suspect, however, that if theirs is a problem of encoding strategy, as suggested above, retardates might well be found attempting to encode all the information of both ears simultaneously by shifting attention from one channel of input to the other. Because of the rapidity of presentation (one pair per half second), though, such a strategy would tend to result in a net loss of such information to recall.

Finally, several studies (Hermelin & O'Connor, 1964; O'Connor & Hermelin, 1965) have shown that, though the digit spans of retardates and normals differ little, a faster rate of decay of STM occurs in the retarded than in normal *S*s. Where does this difference lie? Is it primarily due to decay in the S system as with the senile (cf. Inglis & Sanderson, 1961), or is there also a difference in capacity of the P system? The experiments which follow have been designed to measure both differences in strategy of attention and recall, and in short-term storage capacity between normal and mentally retarded *S*s.

Experiment I

As a preliminary step in testing the hypotheses outlined above, a study was carried out to ensure that the dichotic listening technique would be applicable to the mentally retarded.

Method

Subjects

Three groups of *S*s matched in mental age were used—two mentally retarded and one normal control. The mentally retarded *S*s, obtained from Linekona School for Retarded Children in Honolulu, were grouped into those who were retarded due to organic causes, as determined by medical report (Group O), and those presumably retarded for cultural-familial reasons—at least with no known organic cause (Group F). Normal *S*s (Group N) were obtained from the University Elementary School. The *S*s in these groups ranged in mental age (MA) from 8 years 3 months to 11 years 6 months with the mean IQ for each group, as measured by the Wechsler Intelligence Scale for Children (WISC) as follows: 70.8 for Group O; 68.8 for Group F; and 110.8 for Group N. No *S* showed any impairment of hearing.

Procedure

The apparatus used to administer the binaural stimuli consisted of a Sony two-channel tape recorder (Model 464CS) played into a pair of Sharpe headphones (Model HA10). Different sets of digits, taken from Inglis and Caird (1963) were recorded on each channel as shown in Table 1. Each series was recorded so that two numbers, one from each channel, were simultaneously heard by *S*. The digit pairs within each series were recorded at the rate of one pair every one-half second. Care was taken to

TABLE 1

Digits Used for Binaural Stimulation in Experiments I and II

	Channel 1	Channel 2
Practice series		
A	3	Blank
B	Blank	7
C	3	7
Test series	5	8
	7	6
	4	1
	6	3
	39	72
	85	17
	38	59
	65	28
	592	174
	793	462
	479	836
	584	719
	5638	2941
	9754	8362
	6542	7918
	9356	4271
	81342	96571
	74682	31579
	57841	29356
	38671	15429

control the numbers on each channel for timing and intensity. The headphones covering *Ss'* ears were equipped so that each ear received only the digits from one of the two channels.

Each *S*, on first arriving, was seated at a table opposite the experimenter (*E*). The *S* was briefly introduced to the use of the headphones, and then, via headphones, was instructed following Inglis and Caird instructions: "Now listen carefully. You are going to hear a number. I want you to tell me what number you hear." Practice Series A (the spoken Digit 3 on Channel 1) was then played. If *S* responded correctly, the procedure was repeated with Series B. If *S* failed to respond or gave the wrong number, the volume was increased until the correct response was made. Each *S* was then told: "Now you are going to hear two numbers together, one from each ear. Tell me what numbers you hear." The two channels then played the spoken digits 7 and 3 simultaneously (Series C). If *S* responded with the correct digits (i.e., 73 or 37) then the test series were commenced. This procedure provided a practice series allowing *S* time to become used to the experimental situation, and ensuring that group differences were not due to differences in sensory acuity. When *S* was fully acquainted with the procedure, the test series was begun in the order shown. Th longest series presented to *Ss* of this study was the four-pair series shown in Table 1. The *Ss* were informed of each

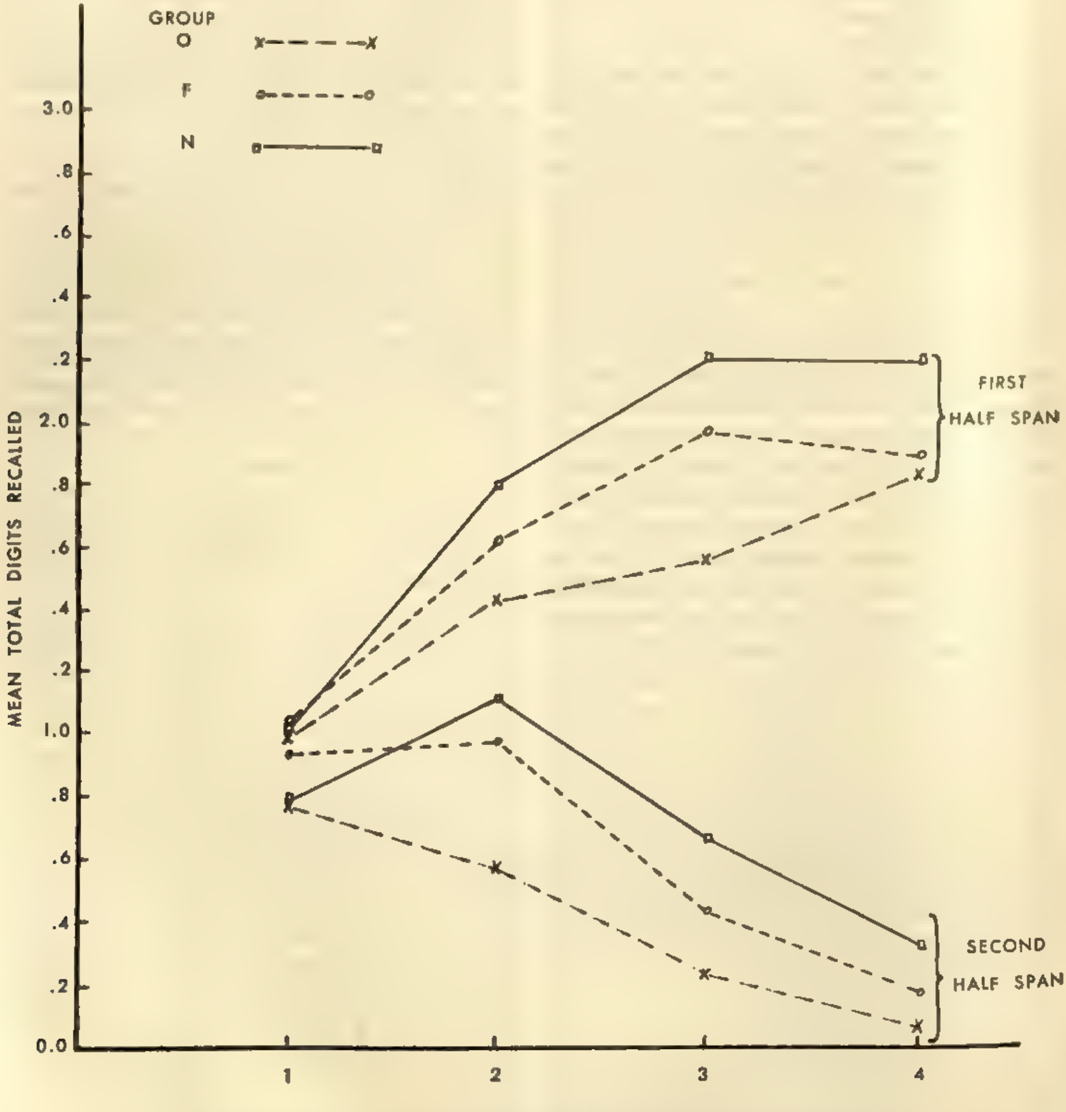

FIG. 1. Mean number of digits recalled per trial for series varying in length.

change in series length by the instruction: "Now you are going to hear [N] numbers, [N/2] in each ear" (where N was 2, 4, 6, or 8). "Tell me what numbers you hear." Between each of the items within a series *S*s were asked, "Now what numbers do you hear?" The *E* recorded *S*'s output on mimeographed score-sheets for later scoring.

Responses were scored following the procedure used by Broadbent (1954, 1956, 1957) and by Inglis and Sanderson (1961). The first digit repeated by *S* determined in each case which channel was taken to be the half-span recalled first. The score obtained was the average number of correct responses for each half-set of digits, taking each digit's position in the series into account.

Results and Discussion

Figure 1 illustrates that, using the dichotic listening technique, group differences in STM are distinguishable. Most striking to cursory examination is that for all groups the digit half-span recalled first is much superior than that recalled second. This result agrees with those obtained by Broadbent (1954) and by Inglis and Sanderson (1961), and concurs with the P and S models advanced by Broadbent (1958). The dichotic technique thus appears to be suitable for more detailed study of group differences between retardates and normals.

The results were subjected to an arcsin transformation (cf. Snedecor, 1956) in order to obtain homogeneity of variance. A Lindquist Type I (Lindquist, 1953) analysis of variance for between-group effects approached significance ($F = 2.77$, $.05 < p < .1$), and thus was suggestive that with better control these groups would in fact differ significantly. One obvious artifact affecting these results was that no attempt had been made to match the groups on the ordinary digit span. Furthermore, Inglis and Caird (1963) have shown a significant effect of CA on immediate memory. Since the retardates were older (mean CA of Group O = 13 years 6 months) and hence had more experience than the normals (mean CA = 7 years 11 months), this may have in effect narrowed such group differences as in fact may be present. In view of such confounding effects of CA on performance Denny (1964) has suggested that two normal control groups be used—an MA and a CA control. Further research should, then, not only match the groups on digit span, but also use the dual control suggested by Denny.

Experiment II

Experiment I presented evidence in support of the notion that both strategy and storage capacity play an important part in differences between normal and mentally retarded *S*s. The purpose of this experiment was to extend and refine that evidence by: studying the problem in more detail, and improving on its experimental design by (*a*) using a dual control as suggested by Denny (1964) and (*b*) matching the experimental and control groups on digit span.

Method

Subjects

Four groups[5]—two mentally retarded and two normal—of 15 *S*s each were used. On the basis of medical evidence available in files on each *S*, 15 clearly organic retardates (9 males and 6 females) were selected, ranging in WISC IQ from 53 to 79. Fifteen familial retardates were then selected, matched in MA and digit span, 1 for each Group O *S*. The *S*s were matched in MA, but *E* kept the IQ and hence the CA of the matchings quite close as well. Plus or minus 3 months MA was considered an adequate match (see Table 2 for a summary of matching data). Items and instructions from the WISC digit span forward were utilized in obtaining a measure of each *S*'s digit span.

The normal control groups were matched as follows: Each of 15 *S*s of the first group (NMA) was matched in MA and digit span with one of the O-group *S*s following the same procedure used in the organic-familial matching above. It should be noted that the only intelligence ratings available for these *S*s were from the California Test for Mental Maturity (CTMM) regularly administered to local elementary school children. It was felt, however, that though not perfect, this rating was adequate as an estimation of normalcy for the MA control group. Each *S* of the second control group (NCA) was similarly matched with one of the O-group *S*s, but in terms of CA rather than MA, keeping their IQ range within the normal range of plus or minus 1 standard deviation from the test mean.

---

[5] Retarded *S*s were obtained from special classes in Kauluwela, Liliuokalani, and Nuuanu Elementary Schools, and from Linekona School. The younger normal *S*s were obtained from Liliuokalani Elementary School, and the older *S*s from Hawaii Baptist Academy. The author expresses his appreciation to the principals of these Honolulu schools, and to Jerry Cochrane, Director of Special Education, Department of Education, Honolulu, Hawaii, whose interest paved the way.

TABLE 2
SUMMARY OF DATA ON WHICH GROUPS WERE MATCHED

Group	CA		MA		IQ[a]		Digit span	
	Mean	SD	Mean	SD	Mean	SD	Mean	SD
O	13.50	2.09	9.25	1.35	69.13	6.24	4.53	.80
F	13.08	2.10	9.25	1.38	71.13	4.14	4.67	.60
NMA	8.83	1.37	9.25	1.40	105.27	6.76	4.73	.44
NCA	13.50	2.01	14.39	2.56	106.86	5.33	5.60	.98

[a] The IQ scores for Groups O and F are derived from the WISC, but the estimates of IQ for Groups NMA and NCA are based on group tests given in the schools.

Although attempted, it was found that as close a match on digit span as found in the previous three groups was not to be obtained here, so that in fact the digit span of NCA is superior to all other groups (e.g., comparing NCA with NMA, $t = 3.36$, $p < .05$). The matching procedure is discussed further below.

Procedure

The apparatus, procedure, and instructions were the same as those used in Experiment I except as follows: (*a*) the items from the WISC digit span forward were recorded and administered via headphones to all *Ss*, for matching purposes, before proceeding with the instructions of the experiment proper. Plus or minus one digit was allowed as acceptable for matching the digit span of individual *Ss* in Groups F and NMA with those of *Ss* in group O; (*b*) whereas the one-pair series contributed little to the study of group differences, this material was dropped from presentation, and the five-pair series shown in Table 1 was added; and, (*c*) the test series were presented in a partially counterbalanced order—half the *Ss* in each group receiving the two-pair material first (the order shown in Table 1), the other half first receiving the five-pair material (reverse order).

Scoring. Two scoring procedures were utilized: (*a*) *ear order*—the procedure described and used in Experiment I. This procedure was used as the best estimate of P- and S-system capacities, following Broadbent (1954), Broadbent and Gregory (1961), and the Inglis (1961) and Inglis and Sanderson (1963) studies, and also as a measure of the degree to which the ear order strategy of recall was in use; (*b*) all other systematic orders. This second procedure allowed for the use of other strategies of recall, such as those delineated by Bryden (1962). Consider for example that *S* has heard 653 on his left ear, and 924 simultaneously on his right. A normal *S* might respond 924653. A response of this nature would be scored as 3 correct for the first half-span recalled and 3 for the second half-span for both scoring procedures as the response follows the "ear order." Suppose, however, that a response is 624953 or 654923, then the "ear order" method of scoring would provide values of 1 and 1 for each half-span (2 and 2 for the latter response) as the first number given is taken as the indicator for the half-span first recalled. It can readily be seen, however, that this *S* actually reported all six digits correctly, but rather than following the ear order he switched from ear to ear in an "attempted ear order." The second scoring procedure would, then, record this latter response as 3 and 3. One other systematic recall pattern would be illustrated by a response of 695234, or of 692534. Both these "temporal" responses would yield a score of 1 and 1 for Scoring I, but 3 and 3 for Scoring II. As with Bryden, most of the responses in the following studies fell into the above three responding orders. The relatively small proportion of systematic responses remaining tended to be a minor variant of one of the above, involving either the repetition of a number, or intrusion of an incorrect number.[6] Scoring II, then, involved the application of that systematic order most closely resembling the given response to that response. Thus, "temporal" and "attempted ear order" responses were as acceptable as "ear order" responses.

RESULTS AND DISCUSSION

Figure 2 presents the mean number of items recalled by each group for each series length as determined by the two scoring procedures. Figure 3 more clearly reveals the overall group differences by combining the data from both half-spans recalled. These same data are presented in Figure 4 as a proportion of the total number of items presented that were correctly recalled. It is readily apparent from the three figures that Group NCA is much superior to any of the other three groups, with NMA falling above, but clustering with Groups F and O, in that order. From these figures the differences between Scoring I and Scoring II are also readily apparent. As Scoring II contains all

[6] A question might be raised as to the effects of chance on the scores assigned to responses. However, as Bryden (1962 p. 293) has pointed out, the likelihood of guessing the correct numbers in one of the systematic orders is extremely low.

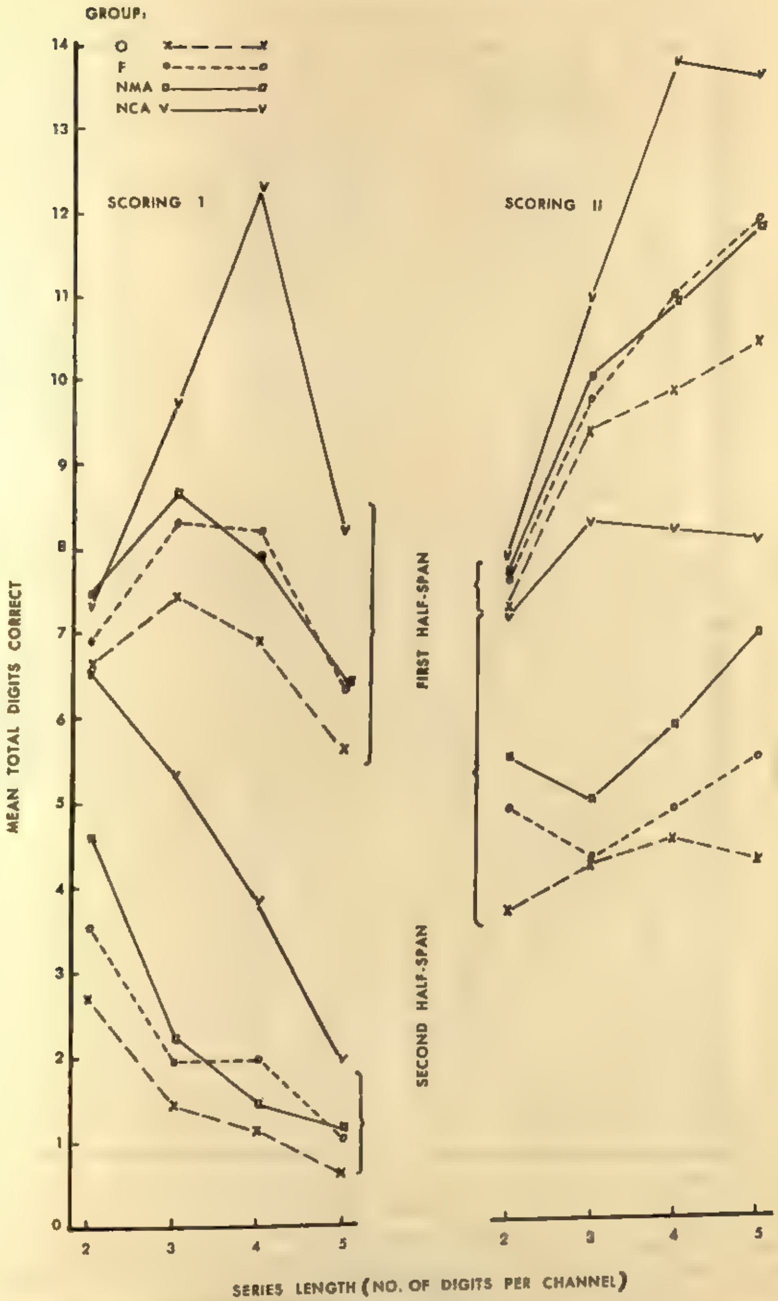

FIG. 2. Amount recalled from series varying in length, summing across the trials of each series length.

the information of Scoring I, plus any additional information retained by strategies of recall other than "ear order," the results of the second scoring are always above those of the first (compare Scoring I and Scoring II on Figures 2, 3, and 4).

After subjecting the original proportions to the standard arcsin transformation (cf. Snedecor, 1956), three Lindquist (1953) Type VI analyses of variance were computed for each scoring of the data. These analyses are summarized in Tables 3 and 4.

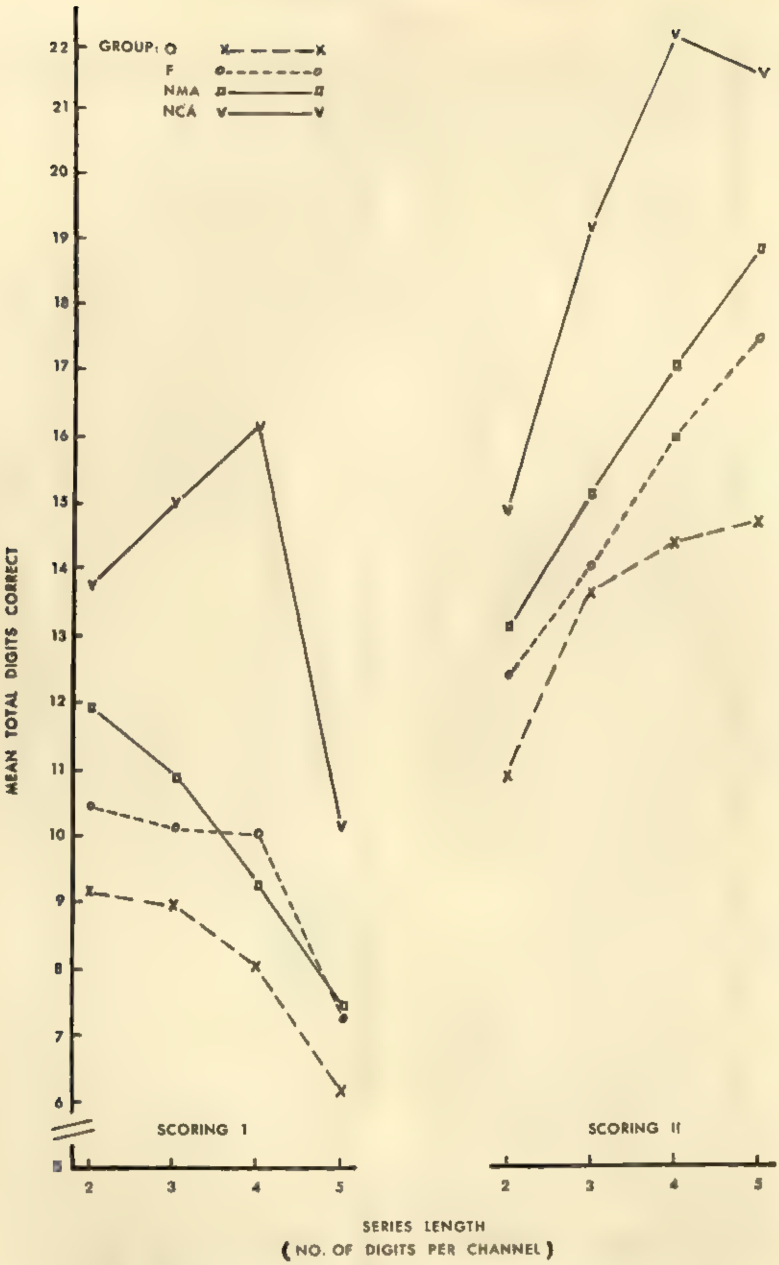

FIG. 3. Total amount recalled in both half-spans from series varying in length, summing across the trials of each series length.

A contrast of Tables 3 and 4 reveals little difference in discriminability by the two scoring procedures used. In both cases the three main effects of Group, Series Length, and Half-span were highly significant on all analyses, with Length × Half-span the only interaction consistently so.

Consider these results in detail. The overall group effect shows the normal *Ss* (both NCA and NMA) to be superior to retardates. Such differences must be considered in the light of the Group × Half-span, Group × Series Length, and the Group × Half-span × Length interactions. The Group × Half-

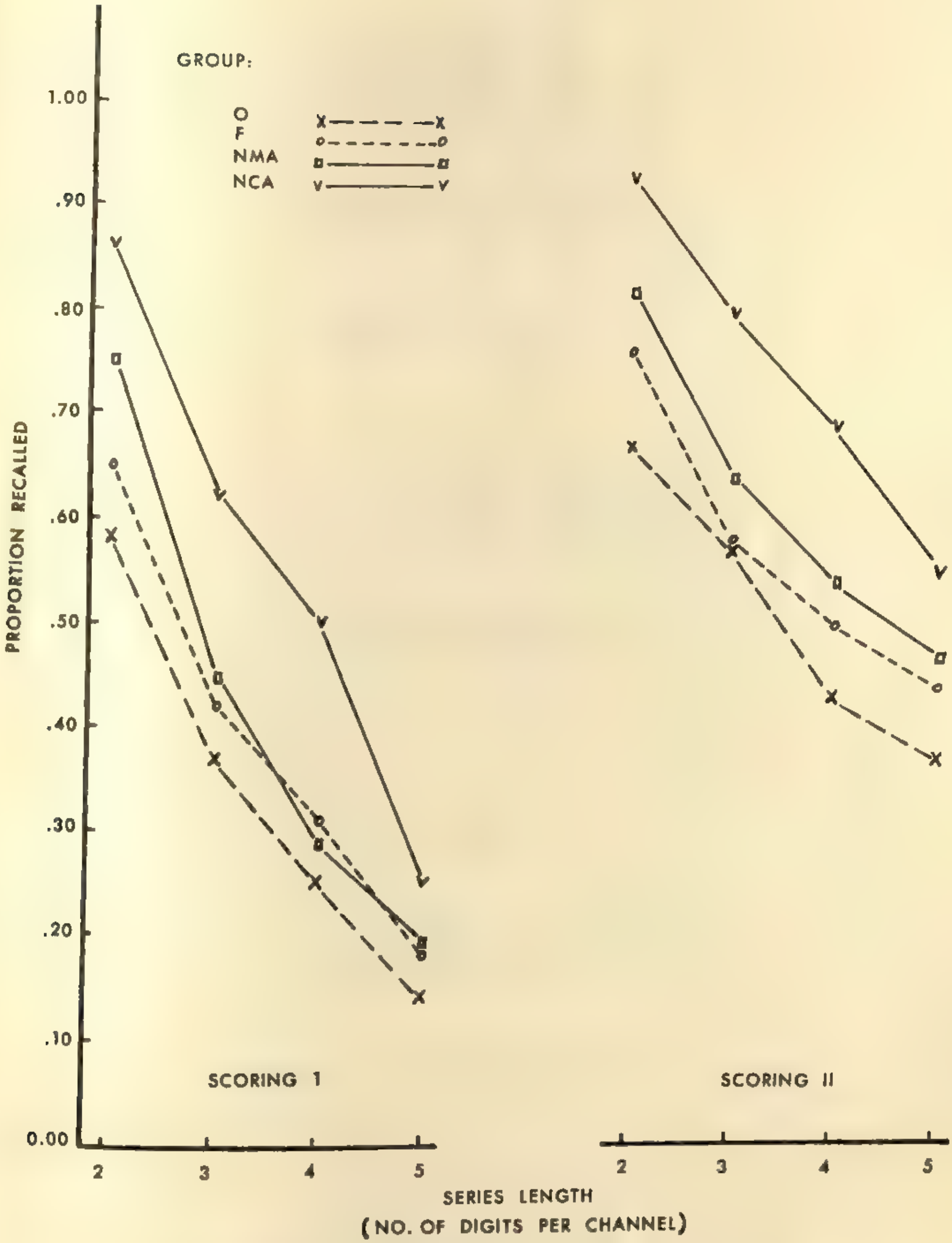

FIG. 4. Proportion of items recalled as a function of series length.

span interaction is of particular interest here because group differences are expected to occur on both halves of information presented. The degree to which the main Group effect is influenced by Half-span and Series Length directly tests the questions about the P and S systems as advanced in the introduction. It can be noted that the Group × Half-span interaction is not significant. This suggests that the main Group effect is

TABLE 3

SUMMARY OF ANALYSES OF VARIANCE, FIRST SCORING OF EXPERIMENT II

Source of Variation	Groups analyzed							
	O × F × NMA			O × F × NCA		NMA × NCA		
	df	MS	F	MS	F	df	MS	F
Between Ss								
Groups (G)	2	16.36	3.25*	106.63	13.60***	1	75.40	10.19**
Error (b)	42	5.06		78.38		28	7.40	
Within Ss								
Series								
Length (L)	3	233.32	131.94***	238.92	161.34***	3	230.59	89.92***
G × L	6	2.90	1.64	3.76	2.54*	3	3.11	1.20
(L × S)w	126	1.77		1.48		84	2.60	
Half-span (H)	1	884.84	204.26***	811.21	156.42***	1	473.66	538.19***
G × H	2	.41	n.s.	3.51	n.s.	1	2.40	2.72
(H × S)w	42	4.33		5.18		28	.88	
L × H	3	11.03	4.52**	8.43	3.09*	3	12.01	5.50**
G × L × H	6	.95	n.s.	7.42	2.72*	3	5.68	2.60
(L × H × S)w	126	2.44		2.73		84	2.18	

* $p < .05$.
** $p < .01$.
*** $p < .001$.

TABLE 4

SUMMARY OF ANALYSES OF VARIANCE, SECOND SCORING OF EXPERIMENT II

Source of Variation	Groups analyzed							
	O × F × NMA			O × F × NCA		NMA × NCA		
	df	*MS*	*F*	*MS*	*F*	*df*	*MS*	*F*
Between *S*s								
Groups (G)	2	19.81	3.69*	105.84	16.64***	1	60.98	15.40***
Error (b)	42	5.37		6.36		28	3.96	
Within *S*s								
Series								
Length (L)	3	120.51	114.08***	129.78	108.53***	3	107.20	87.83***
G × L	6	1.31	1.24	2.29	1.92	3	2.10	1.72
(L × *S*)w	126	1.06		1.20		84	1.22	
Half-span (H)	1	589 68	122.78***	520.70	107.74***	1	235.19	125.18***
G × H	2	2.75	n.s.	9.73	2.01	1	3.22	1.71
(H × *S*)w	42	4.80		4.83		28	1.88	
L × H	3	6.98	3.70*	61.66	31.33***	3	3.26	2.17
G × L × H	6	.75	n.s.	4.17	2.17*	3	4.24	2.82*
(L × H × *S*)w	126	1.89		1.97		84	1.50	

* $p < .05$.
** $p < .01$.
*** $p < .001$.

to be interpreted in terms of differences in both the P and S systems.

The hypothesis with regard to the Group × Length interaction was that group differences should be minimal on the shorter series and increasingly in evidence as series length increased. This interaction, as well as the Group × Length × Half-span triple interaction was found to be significant on the O × F × NCA analysis. The changes representing these interactions can be illustrated as in Figure 5. Using the critical difference technique of determining significant group differences (Lindquist, 1953, pp. 271–272), the above prediction was found to be generally true in the first half-span attended to; that is, no group differences were found on the short series, but the groups did differ significantly on the longer series (see Table 5). In Broadbent's (1958) terms, it would appear the P system (measured by the first half-span) of all groups can ade-

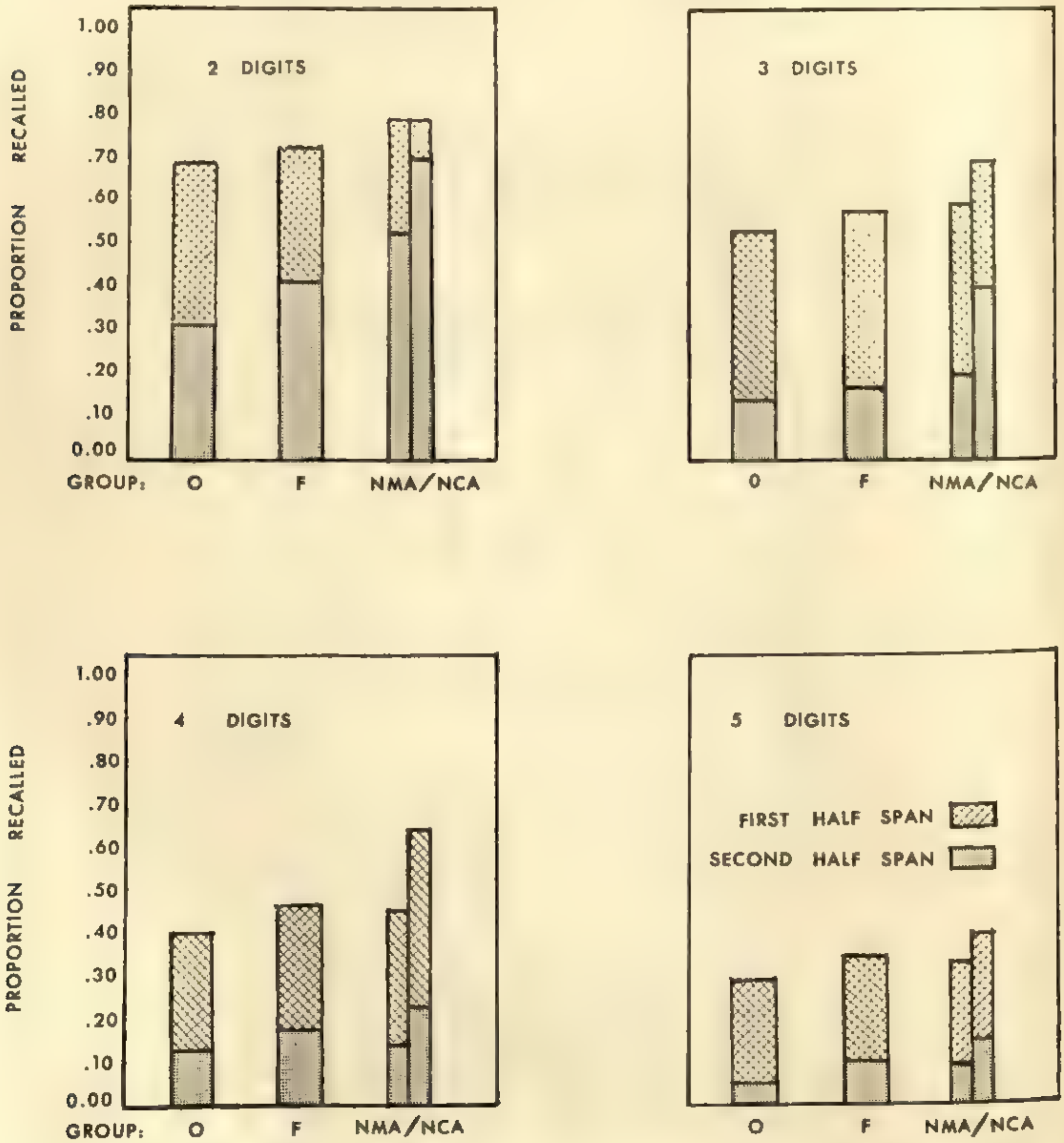

FIG. 5. Group differences in each half-span as a function of series varying in length.

TABLE 5

SIGNIFICANT MEAN GROUP DIFFERENCES FOR BOTH HALF-SPANS, FIRST SCORING OF EXPERIMENT II

Groups compared	Series length in digits per ear				Critical difference for $p < .05$[a]
	2	3	4	5	
First half-span					
O versus F.	n.s.	n.s.	n.s.	n.s.	10.76
O versus NCA	n.s.	16.91	23.63	10.93	10.76
F versus NCA	n.s.	11.39	18.32	n.s.	10.76
Second half-span					
O versus F.	n.s.	n.s.	n.s.	n.s.	14.07
O versus NCA	39.83	25.32	n.s.	n.s.	14.07
F versus NCA	29.21	23.32	n.s.	n.s.	14.07
O versus NMA[b]	21.65				

[a] The critical differences on this table were calculated from the Mean Square within cells obtained from Lindquist Type I analyses carried out separately for each half-span of data (cf. Lindquist, 1953, pp. 271–272). These results are summarized in the Appendix.

[b] All other comparisons of O × F, O × NMA, or F × NMA were not significant.

quately handle a relatively small amount of information. However, whereas the NCA *Ss* are able to increase the load of the P system (up to a point) with increased input of information, the P capacity of retardates appears to remain about the same (see Figure 2) across input conditions.

In the second half-span (second ear attended to) the above prediction did not hold true. In fact, just the reverse was the case. Group NCA had significantly better recall than Group O or F on the shorter series, but such differences diminished to virtually none at the longest (Table 5). It seems, then, that storage capacity of retardates is either uniformly poor in encoding or else subject to rapid decay, the latter being Broadbent's hypothesis. If we agree with Broadbent (1958) and with Broadbent and Gregory (1961) that the first scoring procedure used here is the only valid measure of the S-system, and for the sake of parsimony that the same process occurs in both normals and retardates, then decay would seem to be the correct answer. In Experiment I (Fig. 1, above) it was noticed that the S system of retardates operates about as well as that of normals at the very short series (one digit per ear). With longer series, even as short as two digits per ear (as in this study), recall from the S system falls off and remains low. Similarly, though the short-term storage capacity of normals is relatively good at the shorter series (noted above) as the normal *S* has to attend to an increasingly longer series with his P system, the information in the S system is increasingly subject to decay. Such stimulus decay seems, though, to occur somewhat more slowly in the CA control than in the mentally retarded so that a lack of difference is found between NCA and the retarded groups only at the five-digit series.

Earlier it was noted that the main group effect of the O × F × NMA comparison was significant. This significance was found in both Table 3 and Table 4 above. The suggestion thus was that Group NMA was superior to one or both of the mentally retarded groups. A more detailed analysis of Scoring I with the same three groups, but considering the data from each half-span separately, found that the P system (first half-span) of Group NMA as a whole, though approaching it at points, by parametric analysis never differs significantly from that of either Group F or Group O. The S system of Group NMA, however, is superior, but only to Group O and this on the shortest series alone. A similar analysis of Scoring II presented the same results. Taken in the light of the O × F × NCA comparison above these results suggest that the capacity of the P systems of Groups O, F, and NMA are very similar. Furthermore, these normals of comparable mental age seem to have the same problem of short-term storage decay that the mentally retarded do (this might be taken as supportive evidence for the utility of the concept of

"mental age"). The fact that Group NMA is significantly superior to Group O on the shortest of the series may well indicate that the information decay process in the S system of these normals is somewhat slower than in the retardates, but not quite as slow as that of Group NCA.

Consider next the two main factors of half-span recalled, and series length. All previous evidence (i.e., Broadbent, 1954; Dodwell, 1964; Inglis & Sanderson, 1961) suggested that gross and significant differences would be found between the two half-spans; that is to say, the first channel attended to should be much more accurately recalled than the second. The results of this study corroborate such previous research and thus are supportive of Broadbent's theory in this respect. The significant series length effect is similarly not of particular interest except again insofar as to corroborate previous evidences; namely, that an increase in length of series nets a decrease in percentage of items correctly recalled.

EXPERIMENT III

Earlier it was noted that normal *Ss* will generally use an "ear order" of recall in the dichotic situation with an occasional lapse to some other recall order. This statement holds true primarily for rates of presentation as rapid as one pair per half second. Broadbent (1954) and Bryden (1962) have, however, also shown that at slower rates of presentation, such as one pair per 2 seconds, one of the other recall strategies, primarily recall by temporal order, will more frequently be used; that is, the numbers will tend to be recalled in their order of arrival. As was noted in the introduction, although Broadbent (1958) accounted for this phenomenon in terms of a "switching mechanism," it was felt that this change in recall might better be thought of in terms of recall strategy. At slow rates of presentation *S* can most readily recall the information in the temporal order; whereas, at faster rates of presentation it becomes optimal for the normal *S* to focus attention on all the information presented in one channel first before switching, thus avoiding the wastage of time spent in switching and a concomitant loss of information.

Evidence from both Experiments I and II showed that even at the fastest rate of presentation mentioned above, retardates frequently will use some order of recall other than the ear order. That is, they tend to alternate their attention as the normals (particularly Group NCA) do. This experiment was designed, then, to investigate whether the retardates can be induced to use the more efficient ear order of recall by using a rate of presentation faster than that previously used.

METHOD

Subjects

Dodwell (1964) in a carefully controlled series of studies, found practice effects on this type of task to be virtually nonexistent; thus the same *Ss* were used in this experiment as in Experiment II; namely, two groups of retarded (organic and familial) and two normal control groups (MA and CA controls). There were 15 *Ss* per group.

Procedure

The equipment used in this experiment was the same as that used previously.

Twenty-four three-pair series of numbers were recorded, using numbers from 1 to 10. Each series consisted of six different numbers. Six series were recorded at each of the following rates: one pair per 2 seconds, one pair per second, one pair per half second, and one pair per quarter second. The four conditions were presented in a partially counterbalanced order, half of the *Ss* in each group beginning with the slow rate of presentation, and half with the fast rate.

Scoring. The same two scoring procedures as used in Experiment II were used in this study. However, as the second scoring method is an estimate of over-all correct output, an additional measure was obtained in this study—a difference measure between the two scoring procedures. It was noted in Experiment II, above, that when the ear order of report was used predominantly, the difference between the two scoring procedures was minimal. When, however, some other strategy of recall is utilized by *S*, then the difference in scores obtained by the two procedures increases. The relative usage of the ear order of recall as compared with other strategies, can thus readily be determined. For normal *Ss* this Difference score should be low at the fast rates of presentation and increase at the slower speeds. This would follow from the evidence presented by Broadbent (1954) and Bryden (1962) which shows that normals tend to use the "ear order" of recall at fast rates of presentation, but switch to a temporal order of recall at slower rates.

RESULTS AND DISCUSSION

As was indicated above, the clue as to whether or not retardates can be induced to use the more effective ear order of recall lies in the difference scores. Figure 6 presents the difference scores graphically. The prediction that these scores should be low at the fast rates of presentation and large on the slower speeds was tested by a Lindquist Type I analysis of variance, summarized in Table 6.

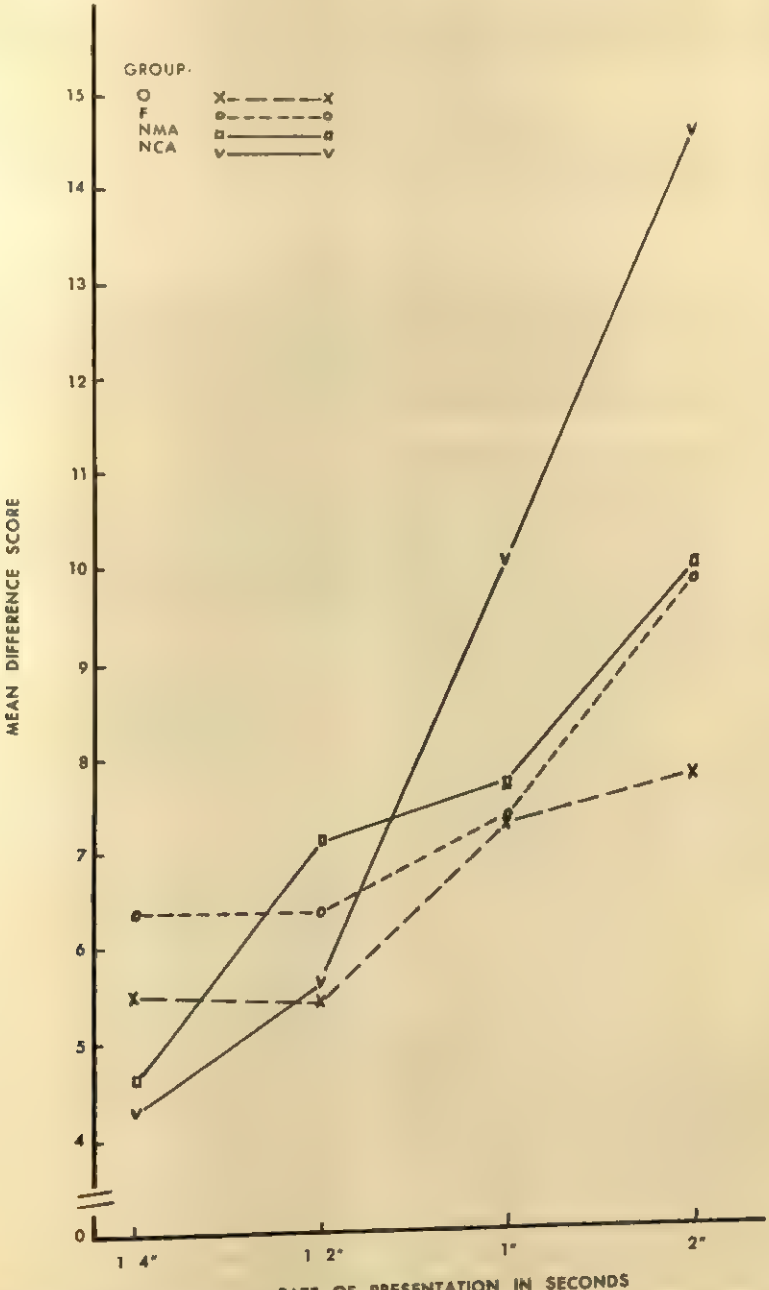

FIG. 6. Group differences in the use of recall strategies as a function of rate of series presentation, calculated by taking the difference between the two scoring procedures used.

TABLE 6

SUMMARY OF ANALYSES OF VARIANCE OF DIFFERENCE SCORES, EXPERIMENT III

Source of variation	Groups analyzed							
	O × F × NCA			O × F × NMA		NMA × NCA		
	df	*MS*	*F*	*MS*	*F*	*df*	*MS*	*F*
Between *Ss*								
Groups (G)	2	75.09	1.16	18.20	n.s.	1	53.34	1.19
Error (b)	42	64.90		55.90		28	44.81	
Within *Ss*								
Presentation Speed (Sp)	3	282.12	26.59***	116.84	11.29***	3	354.89	31.43***
G × Sp	6	60.45	5.70***	9.66	n.s.	3	59.60	5.28***
(Sp × *S*)w	126	10.61		10.35		84	11.29	

*** $p < .001$.

The significant main effect of presentation speed indicates that the groups do indeed change their recall strategy with shift in presentation speed, and a glance at the graph indicates that the shift is in the predicted direction.

The significant Group × Presentation Speed interaction at two of the three levels tested, however, suggested that some groups are more affected by the change in speed than others. Most noteworthy is Group NCA, as can be seen in Figure 6. This group most closely follows the prediction as outlined, showing a very marked shift. In other words, at high speed of presentation this group uses the ear order almost exclusively (as measured by Scoring I), but as the speed of presentation is slowed, Group NCA comes to depend largely on other strategies of recall. Group NMA, though not as markedly, is also strongly affected by presentation speed. A Treatments × Subjects analysis of variance on the NMA data alone revealed the effect of speed as highly significant ($F = 6.78$; $df = 3$, $p < .001$).

Of the two groups of retardates only Group F changes strategy to a significant degree. A Treatments × Subjects analysis of variance obtained $F = 3.20$; $df = 3, 42$; $p < .05$. The treatment differences, however, were significant only at the extreme graphic points ($t = 2.18$, $df = 14$, $p < .05$) and thus have to be accepted with caution. Group O, on the other hand, was only mildly affected by the change in speed (t test between lowest and highest graphic points nets $t = 2.09$, $df = 14, p < .1$).

From these results, then, we can conclude that normal *Ss* tend to be quite flexible in their use of recall strategy. The attempt at inducing the retarded to use the more optimal ear order of recall can be regarded as a modest success only. Far more impressive is the rigidity in the behavior manifested by these two groups. Retardates generally, and particularly those of Group O, tend to show very little inclination toward changing their pattern of recall, even when it is strategic to do so.

EXPERIMENTS IVa AND IVb

The Broadbent (1958) attention hypothesis suggests that when dichotic information is received at a fast pace the best strategy that *S* can adopt—indeed, is almost forced to adopt by the hypothetical switching mechanism—is to "fix his attention on one ear" and perceive the digits presented to that ear at the time they are presented. The *S* holds the stimuli presented to the other ear in the S system and goes back to perceive them later. The digits are, so to speak, lined up in the P system in the order in which they are perceived, and so are most easily recalled in that order. If they were to be recalled in any other order—for example pair by pair—then they must be rearranged, which is difficult just as it is difficult to rearrange an ordinary list of digits and say them backwards.

Yntema and Trask (1963), however, have suggested that recall performance entails more of a search process. They assume, in opposition to Broadbent, that *both* members

of a pair are perceived and stored in memory at the time of presentation. The processor (S) then adopts a search plan (a term taken from Miller, Galanter, & Pribram, 1960), with certain search plans being more readily executed than others. It is, for instance, easier for the search to go forward in terms of presentation order than to go sideways (as in the recall of individual pairs) or backwards. Following this line of reasoning Yntema and Trask suggest that the items are most easily retrieved ear by ear because they have no other characteristic that so neatly divides them into two groups within which the processor may proceed in temporal order. If, however, another prominent set of characteristics or tags were made available to S, the search process should just as readily follow this order as the ear order. Consider the following example:

Left ear	*Right ear*
0	Good
Room	2
5	Coil

Each pair contains both a digit and a word; and the pairs are presented to S at one per half-second. According to Broadbent's (1958) attention hypothesis, S should most readily recall the information from one ear and then the other. Recalling words and then digits should be difficult, from Broadbent's point of view. According to the search hypothesis, however, the S should just as readily, or perhaps more readily, recall the items by type of information as by ear order; perhaps more readily because each item is unambiguously tagged as a word or a digit, but there may at times be a little uncertainty about the side on which it is heard. Evidence found by Yntema and Trask (1963), as well as by Gray and Wedderburn (1960) and by Bryden (1962, 1964), tends to support this latter line of reasoning.

This experiment was designed to test whether the mentally retarded could adopt a given strategy of recall (search process) as readily as normals. Previous experiments (cf. Gray & Wedderburn, 1960; Yntema & Trask, 1963) used familiar words, or word phrases. It was felt, however, that Ss used in this experiment would be more equally familiar with letters of the alphabet than words. With this in mind, 10 letters, A, E, I, O, U, Y, L, M, R, X, were chosen. It can be noted that except for I and Y none of the letters rhyme with another, and that they all are spoken as a single syllable as the digits are. A short study was carried out to ensure the equivalence of these materials. This is described as Experiment IVa, below.

Experiment IVa

Ten mentally retarded Ss, of the familial-cultural variety, naïve with regard to the dichotic listening task, were obtained from Linekona School. Retarded rather than normal Ss were used as it was felt that of the two the retarded Ss should find the letters and numbers least equivalent.

Sixteen three-pair series, of which 8 contain only letters and 8 only digits were randomly arranged and recorded on the stereo tape in the same manner as in the previous experiments. The pairs within a series were presented at the rate of one pair per half second. The instructions used for the practice series were the same as those used in Experiment I.

Results. As in previous experiments, both scoring techniques were used here. For Scoring I the Mean Total number of items recalled was: numbers $= 24.8$, letters $= 20.7$; the t of the difference $= .47$; $df = 9$, and thus not significant. For Scoring II: numbers $= 32.6$, letters $= 29.3$; $t = .19$, $df = 9$, and similarly not significant. It thus seems that though the recall of numbers is slightly better than that of letters, this difference is minimal and at a chance level.

Experiment IVb

The four groups of Ss (O, organic retarded; F, cultural-familial retarded; NMA, normal MA control; and NCA, normal CA control) used in previous experiments were used in this experiment as well, with the exception of two Ss from Group O who, with their matches in the other groups, who were dropped because of inability to maintain attention. Thus, there were 13 Ss in each subject group.

Equipment was the same as that used previously. The dichotic series each consisted of three pairs, a pair being a letter of

TABLE 7
MATERIALS PRESENTED TO SUBJECTS IN EXPERIMENT IVb

	Channel 1	Channel 2
Practice series	4um	a17
	oei	806
	85e	ui6
	3m9	o7r
	4y0	x3u
Test series	034	Loi
	y8r	7i3
	96a	Ly4
	42u	ao8
	eyL	560
	9L8	e6i
	o8a	3m9
	095	4iu
	ux0	72y
	1ar	L69
	786	uyL
	0ry	e94
Practice series	9a5	x24
	eL3	71y
	0ma	i62
Test series	a17	4um
	xLo	816
	8m9	L2a
	e0a	5y9
	xm5	83i
	4e3	i0x
	umL	968
	84o	ru2
	32i	ao7
	Lmr	502
	5xL	i74
	y27	9mx
Practice series	7x1	a9r
	rm4	20e
	9yo	e65
Test series	u8y	4r5
	6ex	i37
	6y4	r8a
	oax	832
	06L	ar9
	3y5	e2i
	423	mrL
	oi4	96r
	6au	L05
	m53	7xr
	xeL	396
	08o	ry4

the alphabet presented to one ear and a digit presented simultaneously to the other (see Table 7). The digits were three different digits (from 1 to 10), and the letters any three of those used in Experiment IVa. The pairs of a series were recorded at half-second intervals and are presented in Table 7.

At the beginning of the session *E* repeated the 10 letters to *S*, indicating that they were the vowels plus four consonants. The *S*s then repeated the letters back to *E*. The *S* was then informed that this experiment, like the previous one, would always have six items, but always contain three numbers and three letters. The letters heard would always be three of those *S* had just learned. The *S* was also instructed to try to say exactly six items after every series, guessing when he could not remember.

Three conditions were used. In the *Pairs* condition *S*s were instructed to report the first pair of items, then the second pair, and then the third. In this condition *E* always illustrated what was wanted by presenting *S* with an example, and then indicating which items belonged together. This continued until *S* understood what was required of him. In the *Sides* condition half of the *S*s in each group were instructed to give the items on the left in temporal order and then the items on the right in temporal order; with the other half left and right were reversed. In the *Types* condition half were instructed to give the digits in temporal order and then the letters in temporal order; with the other half digits and letters were reversed.

Each *S* made 12 trials under each condition. The 12 trials were made in a block and were preceded by 3 or (for the first block) 5 practice trials made under the same condition. Order of conditions was balanced across *S*s within a group. The 12 lists in a block included three of each of the four possible kinds—that is, no crossings (the digits all on one side and letters on the other), a crossing after the first pair, a crossing after the second pair, and two crossings.

RESULTS AND DISCUSSION

An item was counted as correctly recalled only if it was reported in the correct position. Figures 7 and 8 show the results. Lindquist Type I analyses of variance were computed and found the main effect of recall strategy to be highly significant (see Table 8). A further Treatments × Subjects analysis of variance was also calculated for the data within each subject group to test for the relative effects of the three strategies on

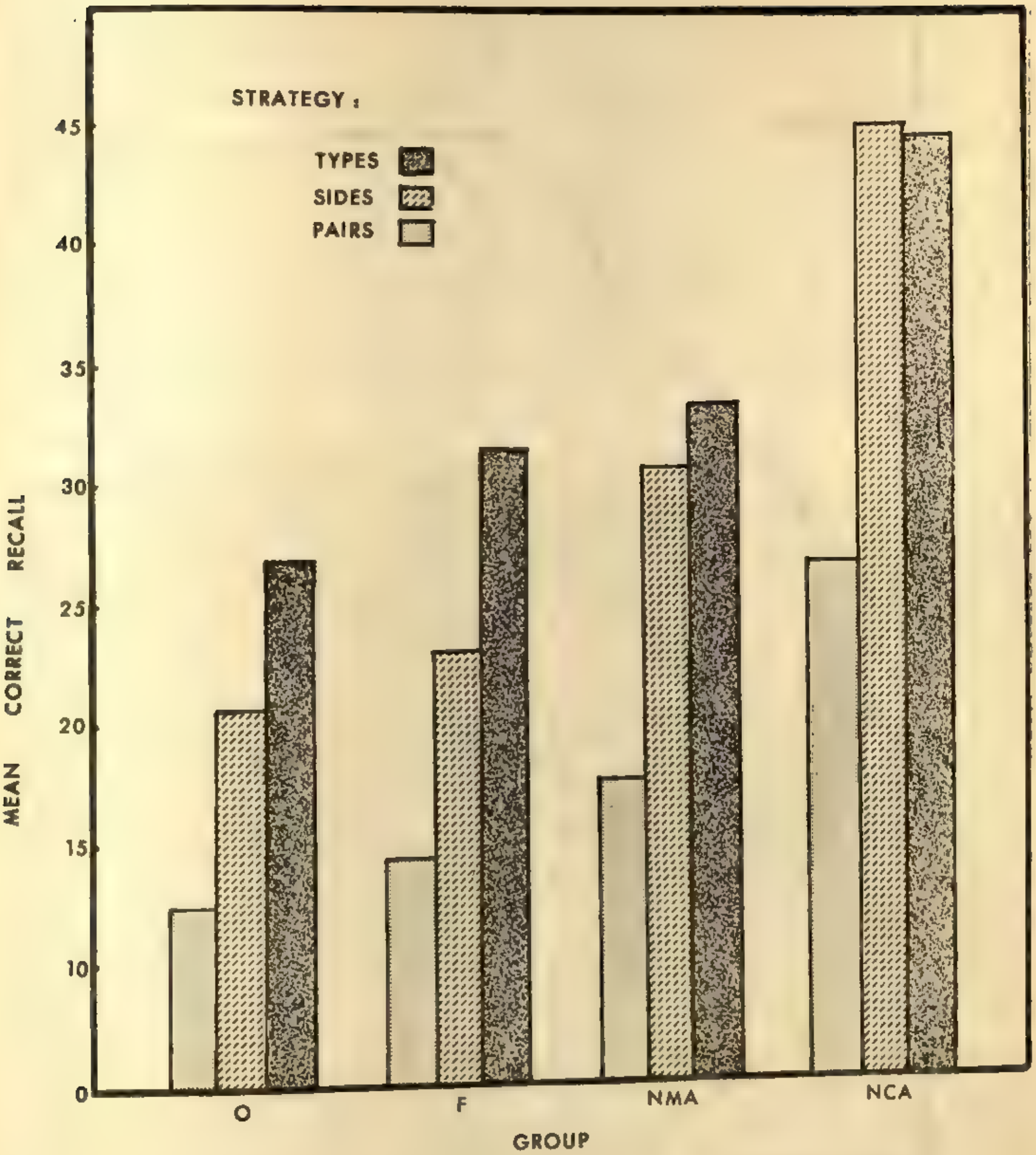

Fig. 7. Differences in recall as a function of strategy, for each subject group.

each group. These analyses similarly found the effect of strategy to be highly significant. Table 9 summarizes significant t-test results as calculated by the critical difference technique from that data. Most noteworthy is that recall is much less accurate when S is instructed to report by simultaneous pairs than when he is instructed to report the items heard on one side and then those heard on the other. This difference was found to be highly significant for all groups.

The crucial comparison with regard to the attention hypothesis is between recall by Types of material and recall by Sides of the head. Under the attention hypothesis recall by Sides of the head should be more accurate. The results obtained here, however, are in agreement with those of Yntema and Trask (1963) who did not find such a difference. In fact, as can readily be seen in Fig-

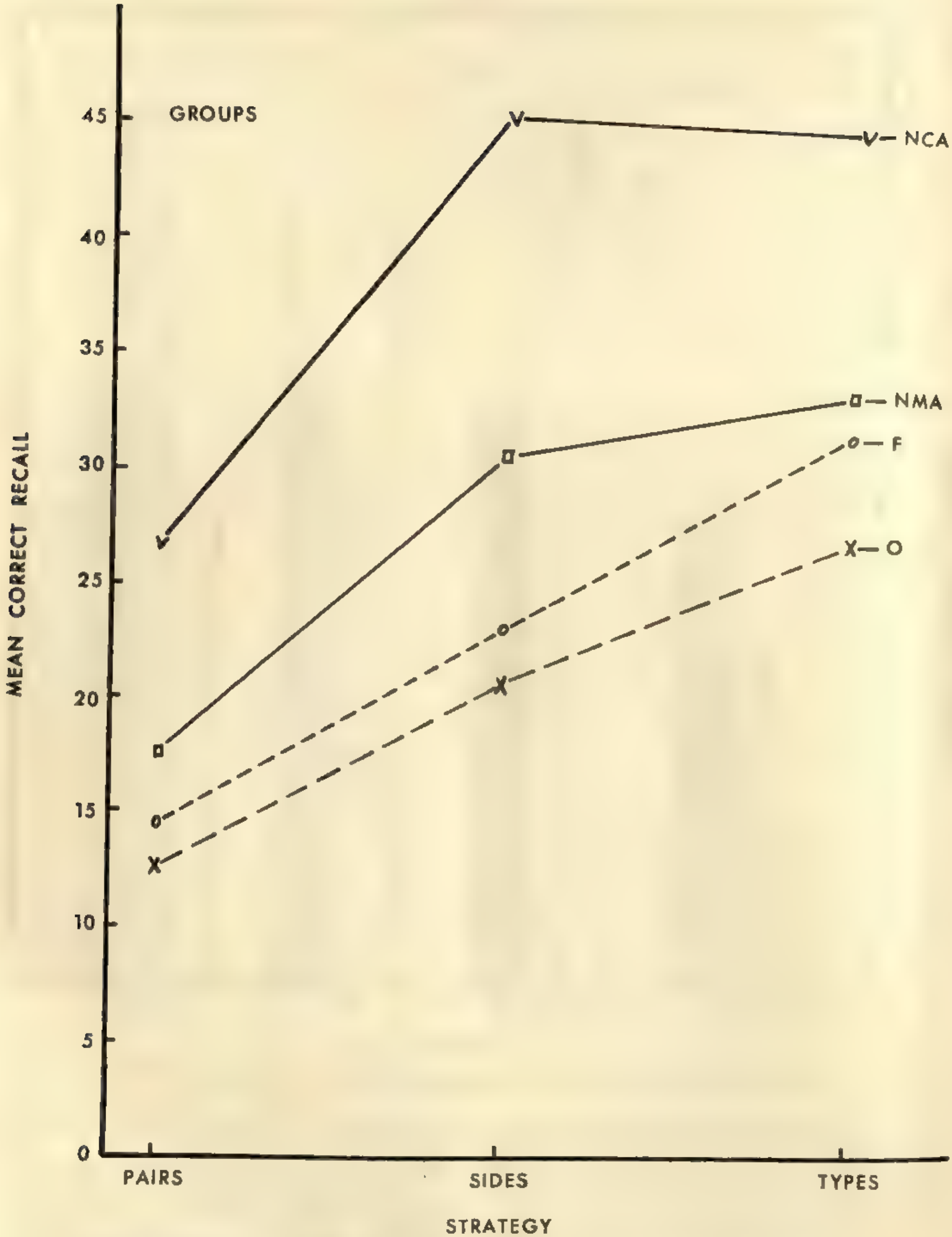

FIG. 8. The same data as in Figure 7, but plotted to more clearly show group differences for each strategy of recall.

ure 7, recall by Types is highly superior in both the retardate groups, while somewhat better but showing no appreciable difference in the normals. This finding supports Yntema's and Trask's conception of the search hypothesis.

It is to be noted, however, that Yntema and Trask found Types of material to be

TABLE 8

LINDQUIST TYPE I ANALYSES OF VARIANCE, EXPERIMENT IVb

Source of variation	O × F × NMA			O × F × NCA		NMA × NCA		
	df	*MS*	*F*	*MS*	*F*	*df*	*MS*	*F*
Between *S*s								
Groups (G)	2	441	1.39	3428	8.01**	1	2269	4.42**
Error (b)	42	317		428		28	513	
Within *S*s								
Strategy	2	2117	40.71***	2449	42.96***	2	2017	51.72**
G × Strategy	4	36	n.s.	113	1.98	2	44	1.10
(Strategy × *S*)w	84	52		57		56	39	

* $p < .05$.
** $p < .01$.
*** $p < .001$.

superior in recall to the Sides condition, while no such difference was found in the normal *S*s tested herein. This difference may well lie in the type of material used for recall. Yntema and Trask utilized words and digits, while this study used letters and digits. As words, of course, are generally high in meaning as compared with letters or digits, it may well be that this difference provided the additional cues to make recall by Types better than that of Sides. Experiment IVa of this paper, however, found that, if anything, recalling letters is slightly more difficult than digits. The failure to find the Types strategy easier than the Sides strategy by the normal *S*s in this experiment then suggests that the cues for materials used here are no more distinctive than those indicating which side of the head the items are heard on.

Although normal *S*s did not distinguish between the Types and Sides conditions the retardates did, finding recall by Types to be easier than recall by Sides. This result might indicate that the Sides condition is somewhat ambiguous—the *S* is not always able to distinguish which stimulus item comes from which side. From the evidence obtained here, it would seem that the mentally retarded are not able to cope with such ambiguity as are the normals.

Of primary interest to this study was the comparison between mentally retarded and normals with regard to their relative ability to handle the various recall strategies. In the overall analyses of variance (Table 8) it was noted that significant group differences occurred only when Group NCA was involved. This effect can be demonstrated by arranging the results as in Figure 8. Table 10, furthermore, demonstrates that Group NCA was superior to all groups on all strategies. In other words, the NCA *S*s were able to utilize even the worst of these strategies (the Pairs condition) reasonably well. Indeed, as can be seen in Figure 8, their poorest performance on the Pairs condition was about as good as the

TABLE 9

STRATEGY DIFFERENCES WITHIN GROUPS

Group	Pairs versus sides	Pairs versus types	Sides versus types	Critical difference for $p < .05$
O	8.16	14.31	n.s.	7.35
F	8.54	16.08	8.54	5.42
NMA	13.07	15.53	n.s.	4.12
NCA	18.62	18.23	n.s.	2.42

TABLE 10

SIGNIFICANT MEAN GROUP DIFFERENCES FOR EACH STRATEGY

Groups compared	Pairs	Sides	Types	Critical difference for $p < .05$
O versus NMA	n.s.	10.07	n.s.	8.51
O versus NCA	14.00	24.46	17.92	9.60
F versus NCA	11.92	22.00	13.07	9.60
NMA versus NCA	n.s.	14.39	11.45	10.04

best performance of the mentally retarded in the Types condition. This again is indicative of the flexibility with which the NCA *Ss* can adopt a given strategy as well as utilize their large P and S capacities.

Both the retarded groups, however, do reasonably well as compared with Group NMA on the Types condition. A result of this nature is probably best interpreted as indicative that both STM capacity and strategy of recall, and the ability to use such strategies is closely linked to an individual's MA (rather than to CA or to IQ), thus also supporting the concept of MA. On the other hand, it should be noted that even though no significant differences exist (by parametric analysis) between Group NMA and the two retarded groups, except between O and NMA on the Sides condition, Groups O and F do fall below the NMA performance on all levels. This might suggest that even when matched in terms of MA the retarded *Ss* are not as readily able to utilize these various strategies.

Discussion and Conclusions

The conjecture of a number of psychologists has been that perhaps the key to learning lies in the understanding of STM, since it seems reasonable to believe that if information cannot pass from STM into permanent storage, learning has not occurred. The experiments above have dealt with the problem of STM, and have been aimed at investigating it at its point of breakdown.

Capacity. Consider the results of the first two experiments in terms of capacity. These experiments presented both the retarded groups (O and F) as well as the normal control groups (NMA and NCA) with dichotic series varying in length from two digits per channel to five per channel (one to four in Experiment I). It has already been noted that in the over-all comparison of retardates with normals (both NMA and NCA), the respective analyses of variance revealed an overall main group effect. This overall effect can be broken down for consideration in terms of the P and S systems.

First of all, the P system of Group NCA, as measured by the first half-set of digits recalled in the Scoring I, was decidedly superior to that of both retarded groups, as well as to Group NMA. As was predicted, this superiority increased with length of the series to be attended to and recalled. At the shortest series length no difference was to be found, but with succeeding series the difference between these groups tended to increase. One could conclude from this that whereas relatively small amount of information can be utilized by the retardates, their P system cannot handle the larger amounts of information. The effective capacity of this system is, then, fairly small for them as compared with normal *Ss* of their same chronological age.

When comparing the P capacity of these retardates with the mental-age control, though, very little difference is found between the groups. The capacity in this system seems about the same for organics, cultural-familial retardates, and normal controls of the same mental age. An examination of the data as plotted in Figure 2 (above) showed that the mean recall for this first half-span of Group O was consistently below that of NMA, but that Groups F and NMA are almost indistinguishable. Although the differences between Groups O and NMA were not statistically significant, the results do suggest that the functional P system of the O group tends to fall below that of NMA. The difference that is in evidence might thus well be due to their relatively greater difficulty in maintaining attention for more than a brief period of time, though this is but post-experimental conjecture. Suffice it to say that the groups did not differ in this respect. Such lack of difference in both the O and especially the F *Ss* suggests that the utility of "mental age" as a concept is meaningful to some degree at least.

What, then, of the S system? It was in this auxiliary storage area that Inglis and Sanderson (1961) found differences between elderly patients with and without memory disorder when such obvious differences could not be distinguished by other STM techniques. It is likewise the system in which Inglis and Caird (1963) later found differences in age groups ranging from 11 through 60 when the groups compared were matched on digit span.

This study similarly found gross differences in the S system of the mentally retarded, especially as compared with the chronological control. But, whereas the P system of NCA was indistinguishable from that of the retarded in the short series and was superior at the long series (as was predicted), the relationship in the S system was found to be just the reverse. That is to say, it was at the shortest series length of Experiment II that the greatest between-group differences occurred. If one can extrapolate evidence from Experiment I to the results of Experiment II, it would be more accurate to say that at the very shortest dichotic series (the one-digit pair) the S system of retardates functions about as well as, though perhaps slightly below, the level of NCA on the average. When, however, the dichotic series increases in length so that more information has to be stored in the S system for longer periods of time, the total output (Fig. 2, Scoring I) as well as the proportional recall falls drastically. As was mentioned in Experiment II (above), this result is highly indicative of rapid storage decay taking place. Such evidence is in keeping with the stimulus-decay model of Broadbent (1958), and also tends to agree with that of Brown (1958). The recall output of Group NCA, though dropping consistently, does not fall as rapidly as that of the retardates, and thus leads one to conclude that the rate of decay of the memory trace in their storage system is slower.

What of normal children of like MA? In the introduction of this paper, it was noted that Hermelin and O'Connor (1964), using a delayed recall type of task, found a faster rate of decay in the STM of imbecils (ranging in IQ from 41 to 54) than a normal MA control group. The evidence in Experiment II suggests that this finding is a result of decay chiefly occurring in the S system. Though the retarded *S*s used in this experiment were considerably superior in intelligence (ranging in IQ from 53 to 79) to those used in the Hermelin and O'Connor study, a difference in decay rate was nevertheless in evidence. Group NMA's recall performance from the second half-span is significantly better than that of Group O at the two-digit series. Whereas most of the imbeciles used by Hermelin and O'Connor were probably of an organic nature, these results would seem to agree with theirs, and furthermore, to pinpoint the decay to the S system. The fact that such statistical significance is not carried over to the longer series of this study indicates that the decay rate of the S system in normal *S*s at this young chronological age is also fairly rapid, though not as rapid as that of the mentally retarded.

Thus far we have considered Experiment II only in terms of STM *capacity*. In agreement with Broadbent and Gregory (1961) the above discussion was, for this purpose, limited to the data as presented by the first scoring procedure. Broadbent would argue that when "errors" occur (which, in some cases, are what this writer, Bryden [1962], and others would consider to be strategies of recall *other* than ear order) one cannot clearly tell how much of the report given can be attributed to the P system and how much to the S system. For this reason, Scoring I is considered by Broadbent to be the more adequate measure of the capacities of these systems. Comparing the statistical analyses of Scoring II with those of Scoring I suggests that the conclusions reached above are not far wrong. The relative relationships between the groups stay the same in both half-spans (compare Tables 3 and 4 above). However, greater informational output of *S*s, as determined by Scoring II, would seem to indicate that Scoring I slightly underestimates the capacities of the two systems. This underestimation would appear to be fairly constant for the four groups, as the difference between Scoring I and Scoring II seems to be about the same at all series lengths (compare Scoring I with Scoring II in Figs. 2, 3, and 4).

Strategy. What, then, about group differences in strategy of recall? Comparing Groups O, F, and NMA on Figure 2 again shows that initially these groups operate near capacity (in the P system) for the first half-span of Scoring I. This would indicate that the ear order of recall is in almost exclusive use at the short series (thus ensuring that the decay found in the S system, as discussed above, is not just an artifact of the strategy utilized by *S*s). There is a slight

increase in recall by use of the ear-order strategy through the three-digit series and then a falling off as the series get longer. However, the total number of digits recalled continues to climb, using Scoring II, even in the longer series. This would seem to suggest a gradual shifting, by these groups, to strategies of recall other than ear order as series length increases.

While the above discussion is true for Groups O, F, and NMA, it is noteworthy that Group NCA continues to utilize the ear-order technique to a large degree up to and including the dichotic series that is four digits in length. This group then suddenly drops the ear order of recall in favor of other recall strategies. The total amount of information recalled in these two (four versus five digits) series lengths remains about the same (though, of course, the proportion of the number presented which are recalled continues to drop), but the amount recalled by the ear-order strategy showed a marked decline. Why this sudden change occurs is not easily answered. As noted above, this change from ear order to alternate strategies occurs in Groups O, F, and NMA, as well as NCA. However, in these groups, the shift was gradual and occurred earlier in the series. One might conjecture that as the STM system becomes increasingly overloaded the *Ss* cease to use the system that served them best in the past and *grasp at any system available to them.* It is, in other words, a shift from an active, organizing strategy of recall to a more or less passive one. One might almost think of this shift as something of a "panic" syndrome, though, of course, no overt panic was manifested by the *Ss* except the occasional remark to the effect that, "Boy, this is getting too hard." That the ear order of recall is both a rational and strategic order of recall is indicated by the fact that as long as it is retained by the *Ss* (of all four groups), the climb in total number of items recalled remained fairly steep with increase in series length—until that point where other recall strategies came into use. At this point the upward trend is flattened, or drops (see both Scorings I and II; on the first half-span in Fig. 2 particularly). Although true for all four groups, this change is most evident for Group NCA.

To summarize, the results of Experiments I and II have revealed that the effective storage capacity of the STM of retardates is much smaller than that of normals of the same chronological age. This is true both with regard to the P and the S system. As compared with normals the same MA, however, retardates are remarkably similar in P-system capacity, but differ somewhat (though not too much) in the S system. The relative differences in P capacity of Groups O, F, and NMA, as compared with NCA, are probably largely a problem of encoding strategy used, although, as indicated below, an inherently smaller capacity cannot be ruled out. In terms of S-system capacity it would seem that there is a real difference between these groups in terms of speed with which information in this system will decay—perhaps a maturational factor.

As was noted above, relatively less use was made of the ear order of recall by Groups O, F, and NMA than by Group NCA, except for the two-digit series. The third experiment was conducted to see whether this situation could be altered by varying the rates of presentation of the stimulus material. The expectation, derived from evidence presented by Broadbent (1954) and Bryden (1962), was that at very rapid presentation rates (i.e., one digit per quarter second) the normal *Ss* would use the ear order of recall amost exclusively, but that at slow rates (i.e., one pair every 2 seconds) this particular strategy would be used very little. As Figure 6 shows, this effect did occur in both normal groups, though most markedly in Group NCA. The retardates, on the other hand, were characterized by a more fixed technique of recall, with Group O manifesting this trait to a more marked degree than Group F. Group O showed no statistical change in strategy from rate to rate. Group F did demonstrate a small but significant change when comparing Difference scores from the fastest with those of the slowest speeds of presentation, thus placing it intermediate in position between Groups O and NMA. What

this evidence tells us is not so much that normals will change their strategy of recall with changes in rate at which information is presented; rather, that normals are flexible enough, when handling information, to search for and utilize better strategies when their previous ones have broken down. The fact that Group NMA does not manifest as marked a shift as that of NCA suggests that this procedure is a matter of learning and practice, although, perhaps, a minimal STM capacity must first be available. The two retarded groups tend to exhibit a very limited amount of such flexibility, suggesting that they tend to assume one strategy and "hang on to it," regardless of whether or not it is strategic to do so.

Experiment IVb was similarly carried out to study strategy of recall in retardates, but was designed to test the ease with which they could adopt a given strategy, as well as contrasting the search hypothesis of Yntema and Trask (1963) with Broadbent's (1958) attention hypothesis. According to the search hypothesis it should be no more difficult, and perhaps somewhat easier to recall dichotic information in terms of types of information heard (i.e., digits and letters) than to report the same information following the ear order of recall. The results revealed that the retardates found it much easier to recall the information in terms of Types rather than to recall all of the information heard in one ear before that heard in the other (Sides). The normals (both NCA and NMA), on the other hand, found these two strategies of recall to be about equal in effectiveness. Both of these results would be in keeping with the search hypothesis.

One might query why the retarded should find it easier to recall by Types than by Sides. One readily available explanation would seem to be that recalling the information in the Sides condition is too ambiguous for retardates. That is to say, in the dichotic situation, especially when the information is fed in via headphones, it is relatively easy to lose track of what information belongs to which ear. For the retardate to keep such information distinct seems to be a difficult strategy to follow. The Types condition, however, presents recall situation in which the items to be distinguished are clearly tagged. Thus, it is this strategy that the retardates (both Group O and Group F) find easiest to handle. Normal *S*s, on the other hand, find the Sides condition no more difficult than that of Types of material. What is a difficult strategy for the retardates can be handled relatively well by both normal groups.

Group differences. The differences discussed above could presumably be due either to learned and/or innate factors. With Group F, a good guess would be that many of these differences could be at least partly attributed to learning factors. Their group name of "cultural-familial" suggests this to be the case. Presumably, children coming from an inadequate cultural environment have not had the opportunity to learn the best of encoding strategies or have not had sufficient experience to allow them to evaluate the relative efficiency of the various strategies. If so, it may be that deliberate training in the use of relatively superior encoding strategies might at least minimize differences between a group such as this and their fellow age-mates. Such specific training would probably have to be begun fairly early, though perhaps early school age (such as the NMA *S*s here) might be adequate. The suggestion of early school age is based on the observation that the development of immediate memory in the NMA *S*s used in these experiments was not too far advanced, as compared to that of normal *S*s of about 10 or 11 years of age (cf. Inglis & Caird, 1963). The determination of what the more sophisticated encoding strategies for the scholastic situation might be must await considerable future research.

As has been seen above, Group F is more like NMA than Group O is. It would seem that Group O suffers from some additional handicap. For them, one cannot be quite so certain that theirs is a problem of learning how to adequately encode information. Rather it might seem, again as their group name (organic) implies, that they may well be deficient in both P and S systems, and that training in strategies would have a lesser effect than with, say, the cultural-

familial type of retardate. Whether or not this is the case also remains for future research, probably at least in part of a physiological nature. Some evidence for such an hypothesis has been advanced by Ellis (1963), who considers retardation to be largely due to stimulus-trace deficits, a position not altogether at odds with the one taken here.

The differences between Groups NMA and NCA are revealing as to how much change occurs between their two respective CAs, for in terms of IQ these two groups were essentially the same. Inglis and Caird (1963) found that the only difference between normals over the age of 11 was essentially in the S system, and, such changes as were in evidence were so only as a trend over a considerable range of ages. Here, however, we found that normals changed a good deal with respect to the apparent capacities of both the S and the P systems, as well as in the ability to utilize the best strategies available, and furthermore, in their flexibility in doing so. There would seem to be both a maturational and a learning effect here. The differences between these two groups in terms of rate of memory storage decay, for instance, might well be maturational in nature, as perhaps is the ability to tolerate large amounts of information and still retain the best strategy. On the other hand such factors as flexibility in the change to more adequate encoding strategies and the search for such strategies might well come about with practice and experience.

On the surface it seems somewhat surprising that the retardates in general, and Group F in particular, did as well as they did as compared to *S*s in the MA control group. Casual experience with both normal and retarded children, of, say, MA 9, would suggest quite marked differences in performance. The normal child appears to be more intelligent in general behavior (despite the equivalent MA) and certainly more adaptable to diverse environmental situations. Furthermore, evidence presented by Jensen (1965) would corroborate such expectations. Perhaps, though, an answer is available in the results of this investigation. Experiment II showed that these groups (O, F, and NMA) had essentially the same P capacity and did not differ much in S capacity. Experiments III and IV, however, revealed that normal *S*s are much more flexible in their use of strategies for information coding and, furthermore, are able to tolerate strategies (such as the Sides condition, above) more ambiguous in nature. The results are, then, perhaps not so surprising after all. The matching of Groups O and F with NMA in terms of MA and digit span suggests that, in terms of potential, these groups are about the same. Evidence from Experiment II corroborates that suggestion. A casual comparison of the overt behavior of these three groups, however, tells us that if their potential is the same, they certainly don't seem to be making the same use of it. Results obtained in Experiments III and IV would tend to corroborate that observation. In terms of STM, at least, these groups seem to have about the same potential, but their use of that potential does differ.

In conclusion, the suggestion that retardates preform so poorly as compared with normals largely because of a limited STM capacity and the inefficient use of encoding strategies would seem to have gained considerable supportive evidence. The studies discussed above demonstrate not only that the effective capacity of mental retardates is smaller than in normals, but also that this limitation may be due, in large part at least, to a lack of flexibility, and hence inefficient strategy usage on the part of the retardate.

SUMMARY

A series of experiments was conducted to investigate STM in mental retardates. The primary purpose of these experiments was to discern whether or not STM capacity and/or strategy of encoding information could account for some of the differences between retardates and normals. This investigation was carried out with the dichotic listening technique is initiated by Broadbent (1958).

An initial study found the usage of this technique to be feasible with retardates. This was followed by three major experiments with four groups of 15 *S*s each. The

groups were as follows: two groups of retardates, one organic (Group O) and one cultural-familial (Group F) in nature, matched in mental age and digit span with a group of normal controls (Group NMA). The fourth group, matched in CA with the two mentally retarded groups, served as a second normal control (Group NCA).

In the first experiment dichotic series of 2, 3, 4, and 5 pairs of numbers were presented to the *Ss* at the rate of one pair every half second. This experiment demonstrated that the effective STM capacity of both retarded groups is much less than that of a comparable CA control, but does not differ greatly from Group NMA. The evidence also indicated that the retardates were subject to a faster rate of information decay in that part of immediate memory which has been termed S system by Broadbent, and is tapped by the second half-set of digits recalled. Comparing the data from the two scoring procedures used, furthermore, suggested that as information-load increased with length of series, *Ss* tended to change in strategy from recalling the digits ear by ear (ear order), to other types of strategies generally less efficient at the rate of presentation used here. Such a change occurred later in the series for Group NCA, than for Groups O, F, or NMA, indicating a greater tolerance for a large information load. This shift appeared to be a change from an "actively organizing" to a "passive" type of recall strategy.

The second experiment held the length of dichotic series constant at three pairs of numbers, but varied the rate of presentation as follows: one pair per quarter second, one pair per half second, one pair per second, and one pair per 2 seconds. This experiment demonstrated a marked degree of flexibility by the normals (both NMA and NCA) in their adaptation of different strategies of recall to the various rates of informational input. Such flexibility was not found in the retardates. At rapid rates of presentation normal *Ss* tended to report the numbers from one ear followed by numbers from the other (as in the first experiment). As the rate slowed, the frequency and accuracy of this order of report decreased while the frequency and accuracy of reporting the material in other orders, such as the order the information arrived at the ears, increased. Such a shift was only partly in evidence in Group F, and not at all in Group O.

The final experiment similarly tested the immediate recall of series six items in length presented two at a time (one to each ear) but held the rate of presentation constant at one pair per half second. In this study, however, each pair of items presented together consisted of a letter of the alphabet and a digit, and the side on which the letter was presented varied haphazardly from pair to pair. For the retarded *Ss* (both Groups O and F) recall was more successful when *S* was instructed to recall the items of one type and then the items of the other type than when instructed to report the items heard on one side and then those heard on the other. Normal *Ss* (NCA and NMA) recalled equally well in both conditions. The conclusion was that, though normals could handle each type of recall strategy equally well, retardates had more difficulty with the greater inherent ambiguity involved in recalling information by sides of the head than by types of material.

In conclusion, the evidence indicated that STM capacity was indeed an important difference between retardates and Group NCA. This deficit in apparent capacity, however, was probably enhanced by the retardates' lack of flexibility in the search for and use of appropriate recall strategies and their manifestation of difficulty with ambiguous types of strategies. Though capacity was essentially the same for Groups O, F, and NMA, the two retarded groups also fell below NMA *Ss* in their ability to adopt a flexible mode of behavior, and to utilize more ambiguous strategies. The differences between Groups NMA and NCA, on the other hand, were indicative of the degree to which both memoric capacity and ability to make use of useful strategies develops in normal individuals over time.

REFERENCES

BERLYNE, D. E. Uncertainty and conflict: A point of contact between information-theory and behavior-theory concepts. *Psychological Review*, 1957, **64**, 329–339.

BROADBENT, D. E. The role of auditory localization in attention and memory span. *Journal of Experimental Psychology,* 1954, **47,** 191–196.

BROADBENT, D. E. Successive responses to simultaneous stimuli. *Quarterly Journal of Psychology,* 1956, **8,** 145–152.

BROADBENT, D. E. Immediate memory and simultaneous stimuli. *Quarterly Journal of Psychology,* 1957, **9,** 1–11.

BROADBENT, D. E. *Perception and communication.* New York: Pergamon Press, 1958.

BROADBENT, D. E., & GREGORY, M. On the recall of stimuli presented alternately to two sense organs. *Quarterly Journal of Psychology,* 1961, **13,** 103–9.

BRUNNER, J. S. Going beyond the information given. *Contemporary approaches to cognition* (Colorado Symposium.) Cambridge: Harvard University Press, 1957, 41–69.

BRYDEN, M. P. Order of report in dichotic listening. *Canadian Journal of Psychology,* 1962, **16,** 291–299.

BRYDEN, M. P. The manipulation of strategies of report in dichotic listening. *Canadian Journal of Psychology,* 1964, **18,** 126–138.

CAIRD, W. K., & INGLIS, J. The short-term storage of auditory and visual two-channel digits by elderly patients with memory disorder. *Journal of Mental Science,* 1961, **107,** 1062–1069.

DEESE, J. *The psychology of learning.* New York: McGraw-Hill, 1958.

DENNY, M. R. Research in learning and performance. In H. A. Stevens & R. Heber (Eds.), *Mental retardation.* Chicago: University of Chicago Press, 1964. Pp. 102–142.

DEUTSCH, J. A. Higher nervous function: The psychological bases of memory. *Annual Review of Physiology,* 1962, **24,** 259–286.

DODWELL, P. C. Some factors affecting the hearing of words presented dichotically. *Canadian Journal of Psychology,* 1964, **18,** 72–91.

ELLIS, N. R. The stimulus trace and behavioral inadequacy. In N. R. Ellis (Ed.), *Handbook of mental deficiency.* New York: McGraw-Hill, 1963. Pp. 134–158.

EMMERICH, D. S., GOLDENBAUM, D. M., HAYDEN, D. L., HOFFMAN, L. S., & TREFFTS, J. L. Meaningfulness as a variable in dichotic hearing. *Journal of Experimental Psychology,* 1965, **69,** 433–436.

GRAY, J. A., & WEDDERBURN, A. A. I. Grouping strategies with simultaneous stimuli. *Quarterly Journal of Psychology,* 1960, **12,** 180–184.

HERMELIN, B., & O'CONNOR, N. Short-term memory in normal and sub-normal children. *American Journal of Mental Deficiency,* 1960, **69,** 121–125.

INGLIS, J. An experimental study of learning and "memory function" in elderly psychiatric patients. *Journal of Mental Science,* 1957, **103,** 796–803.

INGLIS, J. Dichotic stimulation and memory disorder. *Nature,* 1960, **186,** 181–182.

INGLIS, J., & SANDERSON, R. E. Successive responses to simultaneous stimulation in elderly patients with memory disorder. *Journal of Abnormal and Social Psychology,* 1961, **62,** 709–712.

INGLIS, J., & CAIRD, W. K. Age differences in successive responses to simultaneous stimulation. *Canadian Journal of Psychology,* 1963, **17,** 98–105.

JENSEN, A. R. Rote learning in retarded adults and normal children. *American Journal of Mental Deficiency,* 1965, **69,** 828–834.

LINDQUIST, E. F. *Design and analysis of experiments in psychology and education.* Boston: Houghton Mifflin, 1953.

MELTON, A. W. Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 1–21.

MILLER, G. A., GALANTER, E., & PRIBRAM, K. H. minus two: Some limits on our capacity for processing information. *Psychological Review,* 1956, **63,** 81–97.

MILLER, G. A., GALANTER, E., & PRIBMAN, K. H. *Plans and the structure of behavior.* New York: Holt, 1960.

MORAY, N. Broadbent's filter theory: Postulate H and the problem of switching time. *Quarterly Journal of Psychology,* 1960, **12,** 214–220.

NEUFELDT, A. H. The effects of different levels of strategy on the learning of a binary series by "fast" and "slow" learners. Unpublished master's thesis, University of Saskatchewan, 1963.

O'CONNOR, N., & HERMELIN, B. Input restriction and immediate memory decay in normal and subnormal children. *Quarterly Journal of Experimental Psychology,* 1965, **17,** 323–328.

OSBORN, W. J. Associative clustering in organic and familial retardates. *American Journal of Mental Deficiency,* 1960, **65,** 351–357.

OSGOOD, C. E. Method and theory in experimental psychology. New York: Oxford University Press, 1953.

OSGOOD, C. E. A behavioristic analysis of perception and language as cognitive phenomena. In *Contempory approaches to cognition: A behavioristic analysis.* Cambridge, Mass.: Harvard University Press, 1957. Pp. 75–118.

POSNER, M. I. Immediate memory in sequential tasks. *Psychological Bulletin,* 1963, **60,** 333–349.

POSTMAN, L. Short-term memory and incidental learning. In A. W. Melton (Ed.), *Categories of human learning.* New York: Academic Press, 1964. Pp. 138–186.

ROBINSON, H. B., & ROBINSON, N. M. *The mentally retarded child.* New York: McGraw-Hill, 1965.

SNEDECOR, G. W. *Statistical methods.* Ames: Iowa State University Press, 1956.

WEATHERWAX, J., & BENOIT, E. P. Concrete and abstract thinking in organic and non-organic mentally retarded children. *American Journal of Mental Deficiency,* 1957, **62,** 548–553.

WELFORD, A. T. Aging and human skill. New York: Oxford University Press. 1958.

YNTEMA, D. B., & TRASK, F. P. Recall as a search process. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 65–74.

(Received November 17, 1965)

APPENDIX

TABLE A1

Analyses of Variance on Each Half-Span of Data, Experiment I

Source of Variation	Groups analyzed				
	O × F × NCA			O × F × NMA	
	df	*MS*	*F*	*MS*	*F*
First half-span					
Between *Ss*					
Groups (G)	2	35.95	8.23**	6.03	1.46
Error (b)	42	4.36		4.13	
Within *Ss*					
Series					
Length (L)	3	127.65	81.64***	147.09	122.82***
L × G	6	1.99	1.28	.47	n.s.
(L × *S*)w	126	1.56		1.20	
Second half-span					
Between *Ss*					
Groups (G)	2	74.19	8.57***	10.74	2.04
Error (b)	42	8.66		5.26	
Within *Ss*					
Series					
Length (L)	3	119.70	45.26***	97.27	32.31***
L × G	6	9.18	3.47**	3.38	1.12
(L × *S*)w	126	2.64		3.01	

** $p < .01$.
*** $p < .001$.

Vol. 80, No. 13 Whole No. 621, 1966

Psychological Monographs: General and Applied

A REINFORCEMENT ANALYSIS OF GROUP PERFORMANCE[1]

ROBERT GLASER AND DAVID J. KLAUS[2]

American Institutes for Research

2 studies investigated response feedback and reinforcement contingencies occurring in a "team environment." Study I investigated 3-man series teams under conditions of response acquisition, extinction, spontaneous recovery, reacquisition and reextinction. Feedback to team members was based solely on group output. The results suggest team performance can be manipulated using methods which effectively control the behavior of individual organisms. Study II investigated 3-man parallel teams in which a reinforced team response could occur as a function of correct responding by only part of the team. With continued reinforced practice, performance degraded to a level equal to or below initial team performance. These findings are analyzed in terms of an operant conditioning model of team performance.

This article describes an approach to the experimental study of the conditions of training that influence the acquisition and decay of group performance. The general viewpoint taken is one in which group behavior is studied in the same manner in which individual behavior has been investigated successfully in the past. However, it is the behavior of the team, rather than the behavior of its individual members, which is the primary unit of investigation. The approach follows from experimental work which emphasizes the feedback and reinforcement contingencies that are produced as a function of the "group environment" (Glanzer & Glaser, 1961; Klaus & Glaser, 1960). These contingencies are, in turn, a function of (*a*) the probability that appropriate responses will be made by group members and (*b*) the manner in which these individual responses are converted into a collective response.

The kind of group considered here is called a "team." In contrast to a small group, a *team* is highly structured and has relatively formal operating procedures, for example, a submarine crew or a baseball team. A *small group*, on the other hand, is less formal and has few specialized individual tasks, for example, a jury or a committee.

The general characteristics of the analysis used can be illustrated by the example of a two-man team in which a "monitor" obtains information and transmits it to an "operator" who processes the information and transmits the team output. In its simplest form, this output results in a binary contingency, that is, right or wrong, a hit or a miss. The team is arranged in series since both component members must execute a correct response in order for the team to produce a correct response. If the performance of each member is followed by reinforcement only for a correct team product, several predictions can be made about the likelihood of the occurrence of correct individual and team responses under various conditions. When both men are correct, the team response will be followed by reinforcement, and there will be an increase in the probability of correct individual responses. When both men are incorrect, no reinforcement will be forthcoming to either member, and the probability of their incorrect responses will be decreased. When one member responds appropriately and the other does not, the subsequent withholding

[1] These studies were carried out in the Team Training Laboratory at the American Institutes for Research as part of an ongoing research program, Increasing Team Proficiency Through Training, supported by the Office of Naval Research under contract Nonr 2551(00). Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2] The authors wish to acknowledge the contributions of Karl Egerman in conducting Study II and of Jerry Short in reviewing the manuscript.

of reinforcement will result in an extinction trial for the member responding correctly.

In a team situation, reinforcing events which are contingent upon the team response follow the preceding responses of all team members "indiscriminately," that is, every team member is exposed to the same event. For example, in the series-linked team just described, if one member responds incorrectly, no reinforcing feedback is presented to the other member even though he made a correct response. This "confounding"[3] characteristic of team reinforcement suggests one way of defining a team, that is, *a group of individuals who are all reinforced by a single event, the occurrence of which depends on the integrated responding of at least some of the participating members on any one trial.* The emphasis in this concept is that *group* feedback, the reinforcing event, is contingent upon a composite of individual performance.

The major purpose of the studies described in this paper is to assess the feasibility of considering the team as a learning unit which reacts to the presence or absence of reinforcement following a response as do individual organisms observed in a learning laboratory. Accordingly, when the team product is considered as the response, it should exhibit increments or decrements as a function of the properties of the stimulus contingencies following each trial. For example, the team should acquire proficiency in responding when feedback to the team is reinforcing and, once acquired, extinction of the team response should occur if reinforcing feedback is withheld. Study I was designed to test this hypothesis by determining the influence of the presence and absence of group reinforcement on the performance of a series team. Study II considered the more complicated case of a team linked in parallel. In a parallel team a correct response by *either* one *or* more members can produce a correct team response. If team learning can be described in terms of the same principles used to describe individual learning, the study of individual behavior and group behavior fruitfully can share common theoretical structures, similar experimental techniques, and mutual problems.

Methodological Perspective

In studying the performance of a team, one level of analysis is to observe the team as a responding entity. From this point of view, one looks at the stimuli that impinge upon the group and observes the properties of the group responses that occur. It is as if the group were considered an "empty organism" and stimulus-response relationships were observed for study without considering internal mechanisms. The study of team performance on this level can be called "molar" in the sense that the response class under consideration is a group product and not the separate responses of the individual group members. From a molar point of view, group responses can be studied as a function of environmental contingencies without reference to the individuals comprising the team. In contrast, it also would be possible to look at the responses of individual members from a "molecular" point of view. The study of team performance on this level would emphasize individual behavior as it is influenced by the team environment.

Possible relationships that can be studied on each level of investigation and across levels are illustrated in Figure 1. The diagram shows two three-man teams. In Team A, the sequence of response events must occur in much the same way as events in a simple series circuit. In Team B, response events also must occur in series, that is, each man must perform correctly to complete the team task but, in this case, two team members' responses serve as the stimulus inputs to a third member. In Figure 1, the capital letters refer to group stimuli and group responses. The small letters refer to stimuli and responses with respect to individual group members. In Team A, S refers to the stimulus event which initiates group activity. S can be considered as an external stimulus, that is, coming from outside the group's immediate environment. R is the group response which is a function

[3] To use the term suggested by Rosenberg and Hall (1958).

of the performances of the members of the team. S^f is the environmental consequence brought about as a result of the team response. S^f can act as group feedback if it follows the group response. The ovals in Figure 1 refer to individual team members. s_1 refers to the stimulus input for team member 1, r_1 is his response, and s_1^f is the feedback to him or the consequence of his response. S^f and s^f may be different events depending upon the man-machine team arrangement, the remoteness of the individual team member from the occurrence of S^f, and also his opportunity to observe it. The notations in Team B have the same meanings except that S_a and S_b indicate that two independent environmental inputs are fed into this team. (S_a and s_1 and S_b and s_2 may or may not be identical events depending upon the nature of the team task and the construction of the communication arrangement.)

Legitimate variables for study are any of the stimulus and response relationships in Figure 1. For example: the relationship between group input S, group response R, and group feedback S^f; team member response, $r_1, r_2, \ldots r_n$, as a function of group feedback S^f; the relationship between individual feedback s^f and group response R; the relationship between individual feedback s^f and individual response $r_1, r_2, \ldots r_n$. As has been indicated, the relationship between the team member response r and group feedback S^f is especially interesting in situations where the feedback to an individual is not the consequence of his own response but, rather, the consequence of his response as confounded with the responses of other group members.

The two studies described in this paper illustrate applications of this approach in the study of team performance and team learning. Study I emphasizes the relationship between group response R and group feedback S^f, when S^f is considered as a reinforcing stimulus. In this first study, the primary concerns are the observable team output and team feedback events that occur outside the boxes in Figure 1, and the determination of orderly functional relationships between these events. Study II

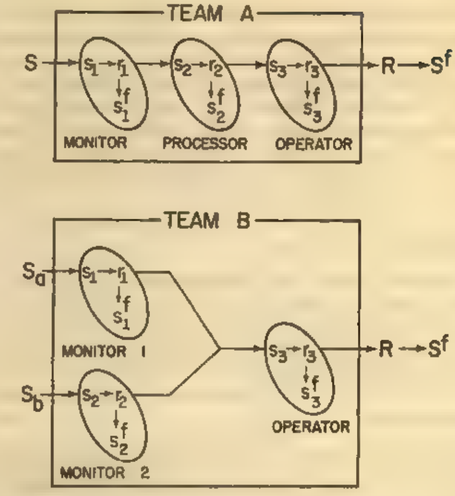

Fig. 1. Reinforcement analysis of series teams.

considers the effect of group feedback S^f on individual team member responses $r_1, r_2, \ldots r_n$ and the subsequent effect on the overall group response R. This study examines a predicted change in R as a function of the learning environment produced when the arrangement of team members permits group reinforcement S^f following inappropriate individual responses. The general purpose of these investigations is to determine the extent to which team learning analyzed at the molar level evidences relationships similar to those that have been identified in studies of individual or single organism learning.

Study I: Acquisition and Extinction of a Team Response

The initial study was designed to consider the team as a unit which responds to the presence or absence of reinforcing events by exhibiting increments or decrements in team performance. The specific hypotheses were that a team response will be learned when group feedback to the team is reinforcing and that extinction of this team response will occur if reinforcing feedback is no longer forthcoming.

Procedure

The units of investigation were three-man teams in which each member was assigned a specific task. Team members were organized in a

series arrangement so that all members were required to perform correctly in order to complete the team task. The task situation was constructed so that no member received any feedback about the accuracy of his own or any other member's performance until the entire team completed the task. When all three members performed correctly, the team as a group received knowledge of a successful trial (S^r).

Subjects

Male high school students were employed as team members. All subjects (*S*s) were at least 16 years old and were selected so that none was in a slow-learning academic group. The *S*s were paid one dollar an hour.

Apparatus

Each *S* was seated in front of a panel on which there was a row of three horizontal lights used to present stimulus displays. Below these lights, the panel contained a counter at *S*'s right and a small red light to his left. A momentary-contact toggle switch by which *S* made a response was centered at the bottom of the panel. From his position at his panel each *S* could see a large counter on the wall, but could not see the other two *S*s.

The response required was the estimation of a time interval, accomplished by either a 2-second or 4-second press of the toggle switch. Tolerance for the 2-second press was $\pm.183$ second and for the 4-second press, $\pm.258$ second, roughly equating the difficulty of the two responses. The *S* made a correct response if he released the switch within the allowable tolerances. The *S*s wore earphones throughout the course of the experiment, and musical broadcasts were played to help mask any stray apparatus noises which might have served as response cues.

A central control apparatus was located in an adjoining room. A 20-pen operations recorder and seven associated five-digit counters recorded stimulus presentations, responses and reinforcements to each *S*, and reinforcements to the team as a whole. Twelve clutch-operated adjustable-duration clocks timed the responses and allowable tolerances for each *S*'s press of the toggle switch.

Method

Six teams of three *S*s each were trained 1½ hours per day (excluding Saturdays, Sundays, and holidays) for a median of 31.5 days (range 21 to 72 days). The experiment consisted of three phases of training: (*a*) Individual Training, (*b*) Stimulus Pattern Training, and (*c*) Team Training. These phases, outlined in Table 1, are described in detail below.

Individual training. The *S*s were randomly assigned to one of the three stations and were called, respectively, Monitor One (M_1), Monitor Two (M_2), and operator (Op). Each *S* was instructed that when he depressed his switch and released it at the proper interval, a point would register on his counter and the adjoining red light would flash for about 1 second. Each daily session during this phase consisted of two half-hour periods during which *S*s practiced each press at their own rate, alternating between the 2-second and 4-second presses every 5 minutes. Daily practice sessions were continued until all three members maintained a proficiency level of .63 or higher for four consecutive 5-minute periods. Proficiency level was calculated by dividing the number of correct presses by the number of total presses for each 5-minute period. After the third member of the team had reached this criterion, four extra 5-minute periods of alternate 2- and 4-second

TABLE 1

DESCRIPTION AND CRITERIA OF THE TRAINING PHASES IN STUDY I

Training phase	Description	Criterion
Individual	*S*s practiced individually on 2- and 4-second time estimations.	Four consecutive 5-minute periods of 63% proficiency or better.
Pattern	Monitors practiced timing responses to light patterns.	None
Three-man team	A. Ten minutes of individual warm-up.	None
	B. All three *S*s performed in a joint effort.	1. Acquisition: four consecutive periods of 10 or more team points. 2. Extinction: four consecutive periods of zero team points. 3. Spontaneous recovery: two consecutive periods of zero team points. 4. Reacquisition: (same as acquisition). 5. Reextinction: (same as extinction).

presses were given in order to provide this last team member with additional practice. The median number of 5-minute periods required for all *S*s to reach criterion was 35 (2.9 hours). The range was 12 periods to 95 periods (1.0 to 7.8 hours), spread over one to eight daily sessions.[4]

Stimulus pattern training. In this phase the monitors learned to discriminate between four light patterns presented on their individual panels. Two of the patterns were used to indicate that the monitors should try a 2-second press; the other two patterns were used to indicate a 4-second press should be attempted. Later, during team training, one of the four light patterns signaled the start of a team trial and indicated whether the monitors should make a 2-second or 4-second response on that particular trial. The pattern discrimination was a simple task, and the monitors learned to perform it adequately in about 10 minutes of self-paced practice during which they responded to 30 to 40 patterns. The sequence of pattern presentation was random except that no identical patterns followed in succession. Since the operator did not have to respond to these stimulus patterns in the team training phase, he did not participate in stimulus pattern training.

Team training. On the first day of the team training phase, *S*s were instructed that all three of them would use their timing skills in working as a three-man team. The team arrangement used in this study is illustrated schematically by Team B in Figure 1. They were told that when all three performed correctly a point would register on the wall counter and a bell would ring. On each trial, the two monitors were presented with one of the four stimulus patterns they had responded to during pattern training. They were instructed that they would not receive any immediate indication about the accuracy of their own responses (the individual panel counter and adjoining red light were made inoperable during team training). The operator's panel displayed two light signals each indicating the duration of a monitor's press. If the operator judged both presses to be accurately executed, he was instructed to make a 4-second press.

If all three team members performed correctly —if the monitors had made an appropriate and accurate response and if the operator had made an accurate 4-second press—a point on the wall counter registered, the bell rang, and a new stimulus pattern automatically was presented to the monitors. All other response combinations were counted as incorrect team responses. When one or both monitors performed incorrectly, the operator's task was to change the stimulus pattern and have the team attempt another point by making a 2-second press. If the operator's response was correct, a new stimulus pattern was presented to the monitors, commencing another trial. If the operator performed incorrectly, no change in any of the stimulus features occurred. When this happened, the operator was instructed to respond again, either with a 2-second or a 4-second press, whichever he judged to be appropriate. As in the case of each monitor, individual feedback was eliminated for the operator. (The extra complication of the operator's task, i.e., judging the monitor's correctness, turned out to be unnecessary and has been omitted whenever appropriate in subsequent data analyses.)

The first 10 minutes of each day during the team training phase was devoted to individual practice as in the individual training phase. On alternate days, the *S*s began either with 5 minutes of practice on the 2-second or 4-second press. During team training, two rest periods were inserted in the daily 1½-hour sessions, one rest period after the first 25 minutes of team training and one after the next 30 minutes.

Team Training Conditions

This first experiment followed an operant conditioning paradigm. In particular, five learning phenomena were selected for detailed examination. As indicated in Table 1, these were: (*a*) team response acquisition, (*b*) team response extinction, (*c*) spontaneous recovery, (*d*) reacquisition, and (*e*) reextinction.

Team response acquisition. During team training all three team members had to perform correctly on any one trial in order for a point to register on the wall counter. The *S*s were trained to a criterion of 10 or more of these points in each of four consecutive 5-minute periods. When the criterion was attained, four additional 5-minute training trials were conducted. Acquisition time for the six teams studied varied considerably. The median number of 5-minute periods to reach criterion for all teams was 53.5 (4.5 hours); the range was from 14 to 402 5-minute periods (1.2 to 33.5 hours).

Team response extinction. Following acquisition to criterion level, the wall counter and bell were made inoperable so that no feedback occurred even though the team continued to perform. A criterion level for extinction was set at zero correct team responses in each of four consecutive 5-minute periods. The median number of 5-minute periods for the six teams to reach this criterion was 56.5 (4.7 hours); the range was from 17 to 157 5-minute periods (1.4 to 13.1 hours). Immediately after a team reached this criterion, it was dismissed from the laboratory for the day. *S*s were instructed to return on the next working day.[5]

---

[4] When an *S* did not reach criterion after 100 5-minute periods, the entire team was dismissed and a new three-man team was obtained. During the course of the experiment four teams were dismissed for failure of one or more of the members to reach individual performance criterion.

[5] At least once during extinction training almost every team remarked to the experimenter that the apparatus must be broken. When this occurred

Spontaneous recovery. Upon returning to the laboratory, and after completing the usual 10-minute warm-up practice on the 2- and 4-second presses, the team again was subjected to extinction. Whether or not the expected increase in response frequency was observed, training was continued to a criterion of zero correct responses in each of two consecutive 5-minute periods. The median number of periods to reach this criterion was 10.5 (0.9 hours); the range was from 3 to 113 5-minute periods (0.3 to 9.4 hours).

Reacquisition. When the criterion for spontaneous recovery was met, the wall counter and the bell were made operable, and the team's correct responses again produced points. The criterion for this phase was identical to the acquisition phase and, as in acquisition, four additional practice periods were conducted. The median number of 5-minute periods to achieve criterion in this condition was 24.5 (2.0 hours); the range was from 4 to 37 5-minute periods (0.3 to 3.1 hours).

Reextinction. Following reacquisition, a second extinction phase was carried out. The criterion for this phase was identical to that used during the extinction phase, four consecutive 5-minute periods each with zero correct responses. The median number of 5-minute periods required for reextinction in five of the six teams was 81 (6.8 hours); the range was from 20 to 200 5-minute periods (1.7 to 16.7 hours).[6]

The Change in Environment between Individual and Team Training

When *S*s were combined into three-man teams, a number of changes in their performance environment occurred. The most significant change was in reinforcement contingencies and ratios. During individual training, each *S* was on a continuous schedule of reinforcement, receiving feedback from the indicators on his own panel. During the team phase, when feedback was changed to the wall counter, *S*s were switched to an aperiodic schedule of reinforcement. Thus, on some occasions a team member might perform correctly but, because of the incorrect performance of another team member, the team did not receive a point. This constituted an extinction trial for the team member who performed correctly. For example, all members of a team might have an individual probability of correct performance of .63 (the minimum performance required of all *S*s before being introduced into the team setting). Since all team members were required to perform correctly on any one trial for a reinforcing event to occur, the probability of a correct team response was at least $(.63)(.63)(.63) = .25$. An *S* who received individual reinforcement for 100% of his correct responses during individual training now received reinforcement only for about 40 per cent of his correct responses under team conditions. The probability of his receiving a team reinforcement when he made a correct individual response depended on the joint probability that the other two team members also were correct on that trial, which would occur $(.63)(.63) = .40$, or 40% of the time. In general, each individual's percent of reinforcement in a series team is a function of the proficiency level of the other team members.

Results: Molar Analysis

Figure 2 shows an individual and a team cumulative performance curve. The dashed line is the performance curve for *one member* of a team during individual response acquisition training. During the last portion of his training, this *S* averaged at least

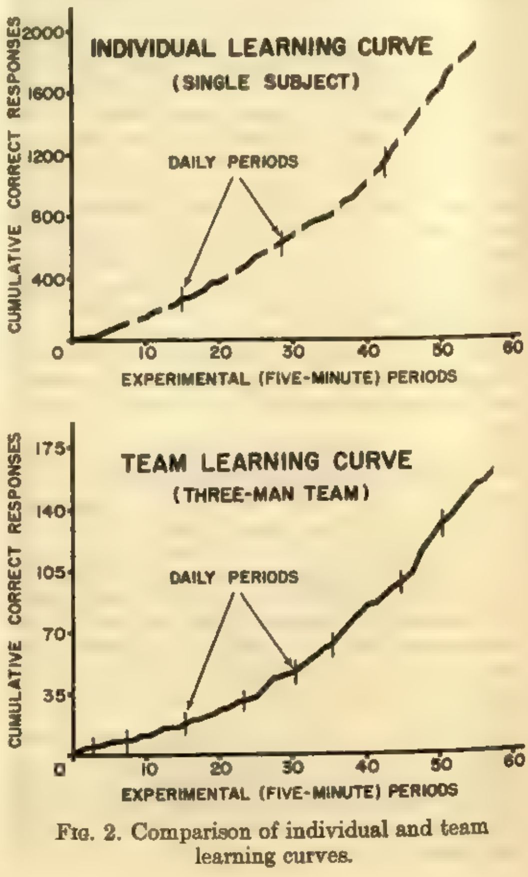

FIG. 2. Comparison of individual and team learning curves.

the experimenter changed switch positions on the central apparatus without being observed, went into the laboratory, and secured a team point by making the appropriate response at each panel. The *S*s apparently were reassured that the apparatus was not malfunctioning.

[6] The one team which did not exhibit reextinction was continued in this condition for a total of 280 5-minute periods (23.3 hours) before being dismissed from the laboratory.

55 correct responses in each 5-minute period. This represents a proficiency level of about 75% correct responses. The solid line shows a *team* response acquisition curve for an early pilot team. This curve is quite similar to the dashed line. However, the rate of responding is less than that shown in the individual curve. Both curves, although representing different rates, show highly similar trends in overall shape and acceleration suggesting that the performance changes that occur during individual and team learning are comparable.

Team Performance Curves

The major results of Study I are presented in the team performance curves shown in Figure 3. These curves are plotted cumulatively, in terms of the number of correct team responses per 5-minute period, through the five training conditions. The number of 5-minute periods required for the teams to reach criterion under each of the experimental conditions is given in Table 2.

Response acquisition. All six teams exhibited learning of the team response during initial response acquisition. For all teams, there was an increase in the number of correct responses indicating that the required task was being learned. On the average, the mean number of correct responses per 5-minute period increased from 2.9 in the first third of the initial acquisition period to 5.7 in the second third and to 8.9 in the last third. There was considerable variability in the shape of the curves and in the time required to reach criterion. The most rapid (Team 5) required only 18 5-minute periods (1.5 hours), while the slowest (Team 4) required 406 periods (33.8 hours).

Response extinction. During extinction, the rate of correct responding for all six teams decreased to zero. This is indicated by the change in slope of the curves. Again, the variability in the time required to reach the extinction criterion was quite marked among the teams. Team 4 extinguished most rapidly, in 17 5-minute periods (1.4 hours), while Team 1 required 157 periods (13.1 hours). For all teams, the extinction of the team response occurred even though 10 minutes of individual warm-up practice under a continuous schedule of reinforcement was provided to all team members at the beginning of each daily session. Apparently, little transfer occurred between individual practice under a continuous reinforcement schedule and team practice on an extinction schedule. Under the first condition, proficiency was maintained but, in the other, proficiency declined over successive trials.

Spontaneous recovery. All teams exhibited this phenomenon, although some more clearly than others. The small number of trials associated with this condition, however, does not permit this phenomenon to be particularly evident in the curves. Teams 2 and 4 each required only three 5-minute periods (15 minutes) to reach criterion for this phase, while Team 3 required 113 periods (9.4 hours).

Reacquisition. Figure 3 shows that with reacquisition training, the teams improved in the rate of correct performance; however, the shape of the curves is not consistent for the six teams. Five of the six teams, all except Team 5, reached criterion in much less time than they did in initial acquisition. This is a typical occurrence when single organisms are retrained under similar conditions. For comparison purposes, the number of trials required for each team to reach acquisition and reacquisition criteria may be seen in Table 2.

Reextinction. Finally, in reextinction, Team 3 failed to achieve the prescribed criterion level. The variability among teams in number of trials to reach this criterion was high. Of the five teams which reached criterion, the fastest team, Team 1, required only 20 5-minute periods (1.7 hours) and the slowest team, Team 4, required 200 periods (16.7 hours). Even after 280 periods (23.3 hours), Team 3 failed to reach criterion. The characteristic single-organism phenomenon of more rapid reextinction than initial extinction was observed in only two teams, Teams 1 and 5.

The changes in team performance illustrated in Figure 3 also are evident when team proficiency, the ratio of the number

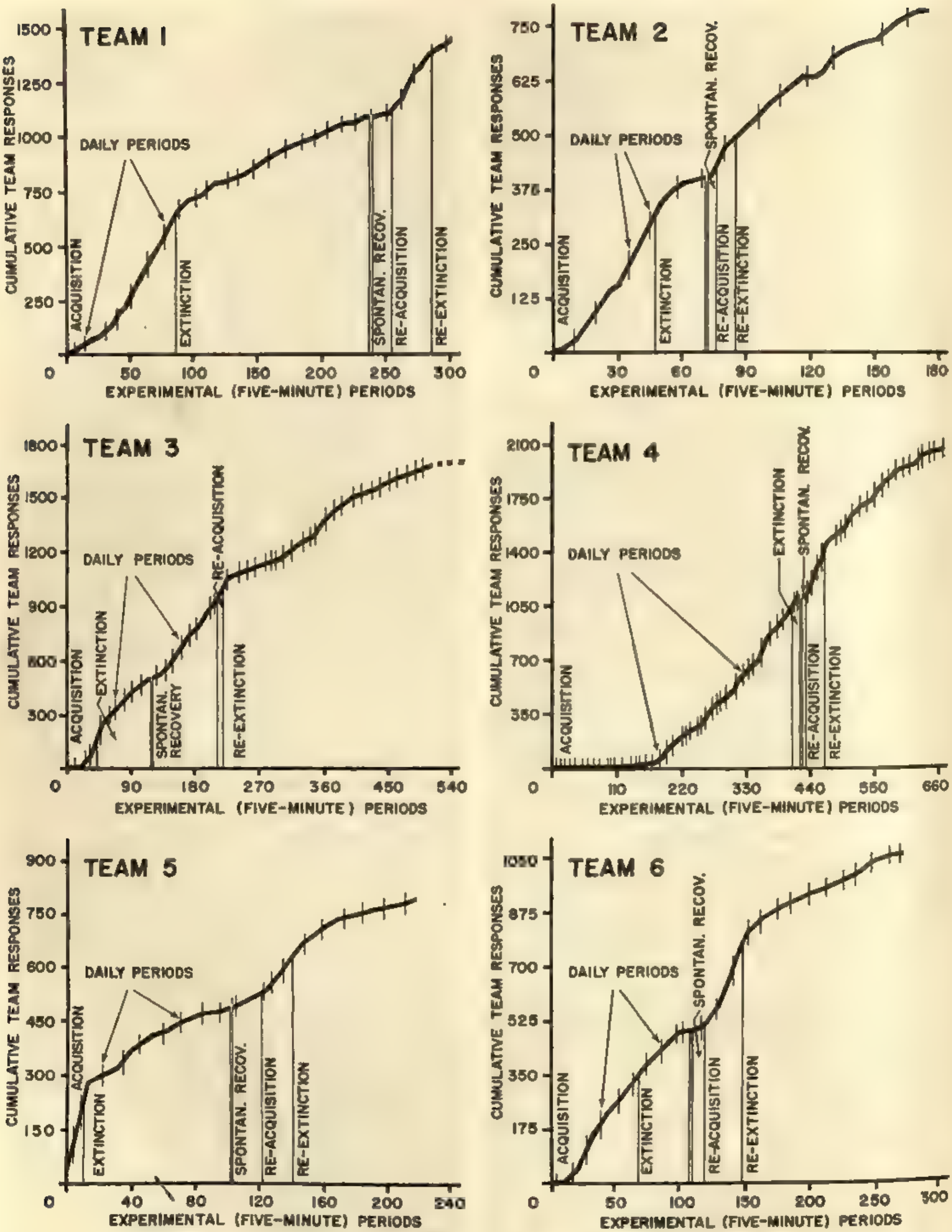

FIG. 3. Cumulative correct team responses for teams in Study I.

of correct team responses to the number of attempts, is compared at various points during the experimental procedure. Table 3 indicates the proficiency of each team, and the median proficiency for all six teams, at the beginning of acquisition, the end of acquisition, the end of extinction, the end of reacquisition, and the end of

reextinction. Data on proficiency during spontaneous recovery are not included because of the very small number of trials required by several of the teams. Proficiency values for the end of acquisition and for reacquisition were obtained on the four postcriterion trials in each of these phases (see procedure) and do not include the criterion trials themselves. Proficiency values for extinction and reextinction were obtained on the last 12 trials preceding the criterion trials; no postcriterion trials were presented following these phases. The proficiency values representing initial proficiency in the team setting were taken from the first 12 trials of this phase. Although there is considerable variability among teams during the 5-minute periods just following or just preceding the criterion trials, the median values given in Table 3 suggest that the influence of having presented or withheld reinforcement is pronounced. There is little likelihood that criterion performance for successive phases was obtained as an artifact of performance variability. Even when the performance of each team is considered sequentially, there is a uniform pattern of rising and falling proficiency as a consequence of the reinforcement contingencies then in effect. It also should be noted that many of the reported proficiencies were obtained by combining records from 2 adjacent days.

The data in Figure 3 and Table 3 indicate that changes in team performance parallel predictions which could be made based on a knowledge of individual organism behavior under similar training conditions. By controlling the occurrence of reinforcement following each team response, it was possible to demonstrate patterns of learning phenomenon usually associated with individual organism performance. The teams showed response acquisition during reinforcement, performance decrement during response extinction, some evidence of spontaneous recovery, and less time to reacquire the response than to acquire it initially. The one dissimilarity between team performance and usual individual organism performance was that reextinction did not necessarily take place more rapidly than did the initial extinction of the team response.

TABLE 2

The Number of 5-Minute Periods of Training for Each Team in Each Experimental Condition

Team	Acquisition[a]	Extinction	Spontaneous recovery	Reacquisition[a]	Reextinction
1	83	157	13	31	20
2	48	24	3	9	91
3	40	72	113	8	280[b]
4	406	17	3	41	200
5	18	88	16	26	76
6	67	41	8	32	122

[a] Includes the four periods given after criterion was reached.

[b] Failed to reach reextinction criterion.

TABLE 3

Study I. Average Proficiency at the End of Individual Training and during Noncriterion Trials under Various Experimental Conditions

Team	Individual training			Predicted team proficiency		Team acquisition		Extinction	Reacquisition	Reextinction
						Initial	Final	Final	Final	Final
	M_1	M_2	Op	($M_1 \times M_2 \times$ Op)	($M_1 \times M_2$)	12 Trials	4 Trials	12 Trials	4 Trials	12 Trials
1	.87	.59	.85	.44	.51	.24	.51	.09	.26	.09
2	.55	.62	.80	.27	.34	.18	.47	.01	.29	.01
3	.70	.68	.74	.35	.48	.01	.65	.16	.54	.10[b]
4	.68	.62	.70	.30	.42	.00	.35	.09	.44	.04
5	.69	.69	.63	.30	.48	.49[a]	.86	.14	.52	.20
6	.73	.68	.71	.35	.50	.04	.16	.07	.78	.07
Mdn.		.69		.33	.48	.10	.49	.08	.49	.08

[a] For all 10 precriterion trials for this team.

[b] Did not reach extinction criterion.

Results: Molecular Analysis

A predicted level of performance for each team was calculated from proficiency measures obtained for each individual team member just prior to beginning team training. This prediction of team proficiency was computed using the multiplicative law of probability; it was assumed that the performance of team members was independent and that team performance, when members were arranged in series, was a function of the probability that each team member would perform correctly on any one trial. For example, if all members of a team had a proficiency of .63 at the end of the individual phase of training, the predicted team proficiency would be (.63) (.63) (.63) = .25. For the monitors this measure of proficiency was a reasonable estimate; for the operator it was an underestimate however, since he could respond more than once in attempting to produce a correct response. This increased opportunity to respond brought the operator's proficiency level close to 1.00 in the team setting so that team proficiency alternately could be predicted by the combined probabilities of the two monitors alone. Table 3 shows the individual proficiencies for M_1, M_2, and Op and predicted team proficiencies using both the two-man and three-man team estimates.

The *Predicted team proficiency* columns of Table 3 contain estimates of each team's performance based upon team member proficiency during the last 12 5-minute periods of individual training. If these estimates are compared to the figures obtained during the first 12 5-minute periods of team training in the *Team acquisition: Initial* column, it is evident that all but one team performed considerably lower than predicted. It is also evident from Table 3 that neither the two-man nor the three-man predictions of team proficiency agree with the initial or final estimates of team performance during acquisition, or show a relationship with the number of trials required by each team to reach the acquisition criterion (Table 2). It might be hypothesized that the performance of a team when it was first formed should reflect the proficiency of its members so that a team composed of high-proficiency members would perform better initially than a team composed of low-proficiency members. It also might be assumed that a team which had high initial proficiency would reach criterion in less trials than one which had low initial proficiency. These hypotheses are central to the approach being examined in this paper, but the data collected for this first study are not appropriate to test either of them. Because of the individual training procedure used, all teams were comprised of members who had been roughly equated in proficiency so that those differences among teams evident in Table 3 largely reflect the random variability which prevented all teams from being as equal in proficiency as had been intended. Thus, no particular correlation between predicted and initial team proficiency was anticipated.

On the other hand, there was a substantial decrease in proficiency for most of the teams as they began performing under team conditions. This difference between observed and predicted team proficiency can be analyzed in terms of the differences in the schedules of reinforcement in the individual and team settings. For the individual phase of training, all *S*s were on a continuous schedule of reinforcement, receiving appropriate feedback following every response. When the three team members each reached at least a .63 level of proficiency, they were combined into a team. At this time, the probability of all three of them performing correctly on any one trial would be about .25. This means that for any one member, only .25/.63 = .40 of his correct responses would be reinforced, representing a considerable drop from the 100% level of reinforcement for every correct response provided during prior individual training. This change in conditions of reinforcement can be hypothesized to be the critical factor in *S*'s performance decrement during the early periods of team training and the resulting poor level of team performance at that time. With subsequent practice, the team members did learn to perform more accurately so that the median

TABLE 4

PERCENTAGE PROFICIENCY VALUES[a] FOR INDIVIDUAL SUBJECTS UNDER CONTINUOUS AND PROGRESSIVE APERIODIC SCHEDULES OF REINFORCEMENT

Subject	Proficiency under continuous reinforcement[b]	Aperiodic schedules							
		15	25	35	40	45	50	55	60
Team 1									
M_1	75	04	12	22		22		31	
M_2	68	11	14	21		34		35	
M_3	43	06	16	23		24		21	
Mean	62	07	14	22		26		29	
Team 2									
M_1	73				21	34	49	53	43
M_2	59				13	20	31	31	36
M_3	65				58	29	35	39	50
Mean	66				31	28	38	41	43

[a] The number of reinforced responses divided by total responses.
[b] Based on the last 12 5-minute periods of individual training.

of the values in the *Team acquisition: Final* column in Table 3 more closely approximates either median value of *Predicted team proficiency*.

If the hypothesis is correct that low initial team performance is a function of the change in the conditions of reinforcement for team members, a similar phenomenon should be observed in a single subject in a nonteam setting under simulated team training conditions. To examine this hypothesis, a supplementary study was undertaken. In this substudy, two teams of three monitors each received the first two training phases, individual and pattern training, as did the members of the six previous teams. However, these teams were treated differently in the team training condition. Instead of working as a team, *S*s were instructed to continue performing as before, making only one press to each light pattern. The panel counters still remained operable; the wall counter was not used. The schedule of reinforcement presented to *S*s as they responded to the light patterns was manipulated over a 5-day acquisition period.

For one team, *S*s received aperiodic reinforcement for correct responses which averaged .15 on Day One, .25 on Day Two, .35 on Day Three, .45 on Day Four, and .55 on Day Five. These values are the proportion of correct responses that were reinforced; for example, if *S* made 100 correct responses on the first day, he would register 15 points distributed randomly over the 100 correct trials. For the second team of three monitors, *S*s had an aperiodic reinforcement schedule of .40, .45, .50, .55, and .60 on consecutive days. For both of these teams, this progressively increasing proportion of reinforcement for each *S* was analogous to the increase that was assumed to occur in the team setting as group performance became more accurate due to the gradual recovery of individual team members from the temporary decrement produced by the change in schedule. The 10-minute warm-up practice given in the main study was provided at the beginning of each day's training.

The results of this substudy are presented in Table 4. The data show that a change from continuous to aperiodic reinforcement does result in a substantial initial decrement in the proficiency of the *S*s. The greater the change between the continuous and aperiodic conditions, the more an individual's performance is likely to decline. A greater initial performance decrement was observed in the three monitors who received the .15 aperiodic schedule following continuous reinforcement than those who received the .40 schedule. With the progressive increase in the proportion of reinforcement, there was an increase in the mean of *S*s' proficiencies except on the second day of the higher schedule.

Conclusions

Study I analyzed team performance at two levels, molar and molecular. In the case of the molar analysis, where only team performance was under consideration, it was observed that learning principles appropriate to single organisms were applicable to a team as a learning entity. At the molecular level, some factors were considered which might explain the observations made at the gross, molar level. In particular, the course of team response acquisition was explained tentatively on the basis of an initial response decrement which was a function of the change in the schedule of reinforcement following the shift from individual to team training. The results of Study I suggest the following:

1. It is feasible to view the team as a single performing entity having response features which are directly affected by team response consequences. Multiman systems appear to demonstrate response acquisition and extinction phenomena characteristically observed in individual organisms. These molar relationships provide a basis for investigating molecular relationships between individual proficiency, team communication arrangements and the consequent courses of team learning and team performance.

2. Although all team members were presented with individual practice at the beginning of each daily session, the teams failed to maintain their performance levels when group feedback was withheld during extinction. Apparently, *S*s who perform in both individual and team conditions can discriminate between the two, and little or no individual proficiency transfers to the team setting. Although subject to further analysis, this hypothesis implies that after being trained to a criterion level by individual training methods, subsequent individual practice by team members may be relatively ineffective in influencing their performance as team members. What probably is more important is practice in which feedback is furnished on the basis of the team product.

3. The initial drop in performance in the team setting suggests that "ease of adjustment" to a group is a function of the change in the schedule of reinforcement to which a member is introduced; this change might be especially marked if the other members of the team are performing at relatively low proficiency levels. In order to minimize the effects of this change, it may be desirable to present individual feedback to team members about their own performance regardless of the team output. It can be hypothesized that under this condition *S*s would not demonstrate the observed decrement when initially placed in the team setting.

STUDY II: EFFECTS OF TEAM REINFORCEMENT IN PARALLEL TEAMS

In contrast to the "series" teams investigated in the first study, Study II considered teams linked in "parallel." In a parallel team, a correct response by any member can produce a correct team response. For a team with a parallel structure, the reinforcing contingencies are more complex than they are for a series team since a team reinforcement can occur even when one or more members have made an *incorrect* response. This potential for reinforcement following incorrect responding is a major property of teams where a group of individuals are all reinforced by a single event, the occurrence of which depends upon the integrated responding of some, but not all, of the members on any one trial.

The defining property of a parallel team is the redundancy which exists among the members that compose the team. Team C in Figure 4 depicts a two-man series team. In Team D in Figure 4 an additional monitor has been added so that the two monitors are in parallel with each other and both are in series with the operator. Whenever one monitor performs correctly, the performance of the other monitor is redundant. (This arrangement typically is employed in performance situations where an incorrect response might have serious consequences. Common sense suggests the overall proportion of incorrect responses will be reduced if more than one team member is assigned to the same task.) If a team is

arranged in this way, a redundant member can make an incorrect response and yet be reinforced along with all of the other members of the team following a correct team response. Over several reinforced trials, these incorrect responses are likely to be strengthened, and, on subsequent occasions, such incorrect behavior would have an increased probability of recurring. Under certain conditions, it is probable that both monitors will exhibit an increase in the frequency of incorrect responses over a series of trials as a function of reinforcement following such responses. Thus, although the addition of an extra team member in parallel with an existing member initially may result in a higher level of team performance, with continued trials the redundant members are likely to show a decrement in performance. This decrement would result in a corresponding decrement in team output and would bring the level of team performance to a point equal to or possibly below that of the previous, nonredundant team.

A decrement in the performance of parallel teams can by hypothesized to be a function of (*a*) the number of trials on which the redundant members were reinforced for incorrect performance, and (*b*) the initial proficiency levels of the members and the relative magnitude of these proficiency levels among the members. Three specific hypotheses can be proposed:

1. With the addition of a redundant member, team output will show an initial increase in the number of correct responses. The size of the increment will depend upon the current proficiencies of the team members: if proficiencies are high when a redundant member is added, increments in team performance will be small. For example, if both the operator (Op) and the monitor (M_1) of a two-man series team have high proficiencies, .86 and .90, respectively, the probability that the team will produce a correct response is (.86) (.90), or .77. If a redundant monitor (M_2) with a proficiency level of .88 is added, the probability of a correct team response is $Op[M_1 + M_2 (1.00 - M_1)]$, or $.86[.90 + .88 (1.00 - .90)]$, which is equal to .85, an increase

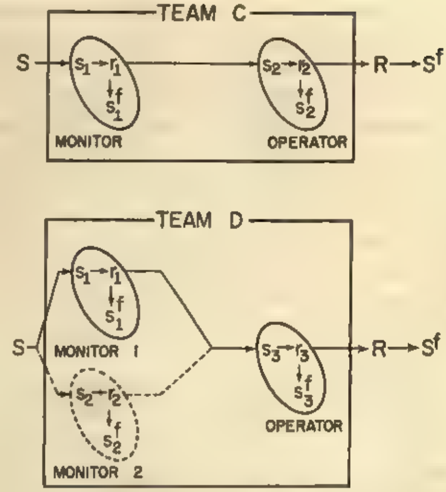

Fig. 4. Reinforcement analysis of team when redundant member is added.

of .08 in overall team proficiency. This is the probability that the operator along with either or both of the monitors will perform correctly on any one trial. On the other hand, if the original two-man team proficiencies are lower, e.g., the operator (Op) at .80 and the first monitor (M_1) at .50, the probability of a correct two-man team response is .40. The addition of a redundant monitor (M_2) with a proficiency level of .88 will result in a parallel team in which the initial probability of a correct team response is .75, an increase of .35 in overall team proficiency. In general, it is hypothesized that when the original two-man team performs at low proficiency levels, the addition of a redundant monitor can add substantially to team output.

2. With subsequent trials, the redundant team will show a performance decrement as a result of reinforcement for incorrect team-member responses. The rate of decrement will be a function of the extent to which a redundant member's incorrect responses are reinforced; this, in turn, will be a function of the proficiency levels of each of the members. In order to explain this further, consider that in a *series* team of two or more members there are three possible response reinforcement contingencies. First, for each of the members of the series team a correct response

can be followed by a team reinforcement; this "appropriate" reinforcement occurs when all members perform correctly. Second, a correct individual response also can be followed by no reinforcement; these extinction trials occur for a team member when he performs correctly but other members perform incorrectly. Third, an incorrect response by a team member in a series team always must be followed by no reinforcement since the team cannot possibly be correct when any member has performed incorrectly.

In a parallel-redundant team, where only one of the monitors along with the operator needs to respond correctly for correct team performance, a fourth reinforcement contingency occurs in addition to the three possibilities which exist for a series team. When a monitor responds incorrectly on the same trial that the other monitor responds correctly, his incorrect response is followed by a team reinforcement. On any one trial, the probability that a particular kind of feedback contingency will occur for a specific team member can be determined on the basis of the proficiency levels of all members of the team. For example, if $M_1 = .70$, $M_2 = .40$, and $Op = .80$, team proficiency (the probability of a correct team response) for any one trial is .66. On this trial, M_1 will be appropriately reinforced when both he and Op respond correctly, which will occur with a probability of (.70) (.80), or .56. For M_2 the probability of team reinforcement following a correct response is .32. The probability of M_1 receiving a team reinforcement following an incorrect response is the probability that M_1 is incorrect (1.00 − .70) on the same trial that M_2 and Op are both correct (.40) (.80); or, (.30) (.40) (.80) = .10. For M_2, the probability of reinforcement following an incorrect response is .34. (A complete analytic description of team learning would need to consider the changing response probabilities of the members from trial to trial instead of just the momentary, one-trial state of affairs considered here.)

The above, hypothetical data show that on trials when M_1 receives a reinforcement, the ratio of correct to incorrect responses is .56 to .10, or approximately 6 to 1. For M_2 the ratio is .32 to .34, or approximately 1 to 1. It is evident that M_1 and M_2, especially M_2, have been placed in an environment in which it may be difficult to maintain the proficiency necessary for successful team performance. As a result of partial reinforcement for incorrect responses, it is possible to hypothesize that both monitors will show an increase in the frequency of incorrect responses and, consequently, that the team will show a performance decrement.

3. If one redundant member's proficiency is very high, no decrement in overall team performance is expected. One redundant member can carry the load for other, parallel members. For example, if M_1 starts with an initially high proficiency and M_2 with a relatively low one, M_2 should show a faster decline in proficiency than M_1. In such a case, it is likely that the ratio of correct to incorrect responses for which M_1 is reinforced will change and approach a more favorable one for maintaining the correct response. Thus, it can be hypothesized that for a team consisting of redundant members who have initially divergent proficiency levels, one very high and the other low, team performance would become primarily a function of the more proficient member and the contribution of the poorer member would become increasingly small. On the other hand, when the divergence in initial proficiency is not great, the performance of both monitors can be expected to deteriorate concurrently (as described in Hypothesis 2 above).

Procedure

Preliminary training. Six parallel teams of three *Ss* each worked for a median of 37 days (range 22 to 54 days). The team tasks assigned to the monitors and operator were the same as in the first study. The experiment consisted of the four phases of training shown in Table 5. Individual training and stimulus pattern training were identical to Study I. For the individual training phase, the median number of 5-minute periods required to reach criterion was 44 (3.7 hours); the range was 35 periods to 88 periods (2.9 to 7.3 hours).

Two-man team training. Each day began with a 10-minute individual warm-up period, as in the

first study. For the remainder of the one and one-half hour session, each monitor worked alternately for 15 minutes with the operator in a two-man series arrangement. During the time that one monitor worked with the operator, the other practiced alone with individual feedback. Both monitors received the same number of practice periods with the operator. Rest periods were inserted, one after the first 25 minutes and one after 55 minutes. The criterion for this phase of the study was four consecutive 5-minute periods during which 15 or more team points were scored by each two-man team. The median number of 5-minute trials to reach this criterion was 37 (3.1 hours), and the range was from 9 to 45 5-minute periods (0.7 hours to 3.7 hours).

Parallel team training. The *S*s were instructed that they would work as a three-man team to score points on the wall counter. The monitors were instructed to respond exactly as they had been doing, that is, with either a 2- or 4-second press for each light pattern. The operator was instructed to observe his panel lights. He was to respond with a 4-second press if he felt either or both of the monitors was correct and with a 2-second press if neither of them was correct. The apparatus was adapted so that only one of the monitors and the operator needed to perform correctly in order for the wall counter and the bell to operate. A correct 4-second press by the operator scored a group point if either or both of the monitors was correct while his 2-second press advanced the program, presenting a new pattern to the monitors. The operator was permitted additional responses until either a team point was scored or until the stimulus pattern was advanced. If the team scored a point, the stimulus pattern was advanced automatically. The first 10 minutes of each session were spent in individual practice of the 2- and 4-second presses as before.

TABLE 5

Description and Criteria of the Training Phases in Study II

Training phase	Description	Criterion
Individual	Same as Study I	Same as Study I
Pattern	Same as Study I	Same as Study I
Two-man team	Monitor One and Operator, and Monitor Two and Operator of each team practiced as a two-man series team.	Four consecutive 5-minute periods each with 15 or more team points.
Redundancy team	The three subjects worked as a three-man team with parallel monitors.	Four consecutive 5-minute periods each with 10 or fewer team points.

Since it was hypothesized that teams in a redundant arrangement would show a decrement with continued performance, the criterion for this phase was four consecutive 5-minute periods of reinforced team performance with 10 or *fewer* group points in each. The *S*s were male high school students at least 16 years old who were paid one dollar an hour. The display-response panels and control apparatus were the same as that employed in the first study.

Results: Molar Analysis

Figure 5 shows the performance curves for the six teams plotted in terms of percentage proficiency per day (the number of correct team responses over the number of total team responses) for both the two-man teams and the subsequent three-man parallel teams. The median number of 5-minute periods for five of the six teams to reach the three-man team criterion was 186 (15.5 hours), with a range from 111 to 437 periods (about 9.2 to 36.3 hours). One team failed to reach criterion in 444 5-minute periods (37 hours), after which the experiment was terminated for that team.

The proficiency of the two-man teams during the 12 periods preceding the criterion trials at the end of two-man team training and the proficiency of the three-man teams during both the first 12 periods at the beginning and the 12 last periods preceding the criterion trials at the end of the parallel team training are given in Table 6. As can be seen from the table, five of the teams demonstrated some increment in performance over that of either two-man component following the initial addition of a redundant team member. As a result of continued training, the proficiency of four of the six teams fell to below the level of performance observed when the three-man teams were first formed. The criterion for ending two-man team training required highly proficient performance, four consecutive 5-minute periods each with 15 or more team points, and the criterion for ending redundancy training required a substantial decrease in proficiency level, to four consecutive 5-minute periods each with 10 or fewer team points (see Table 5). Because the termination of these phases of

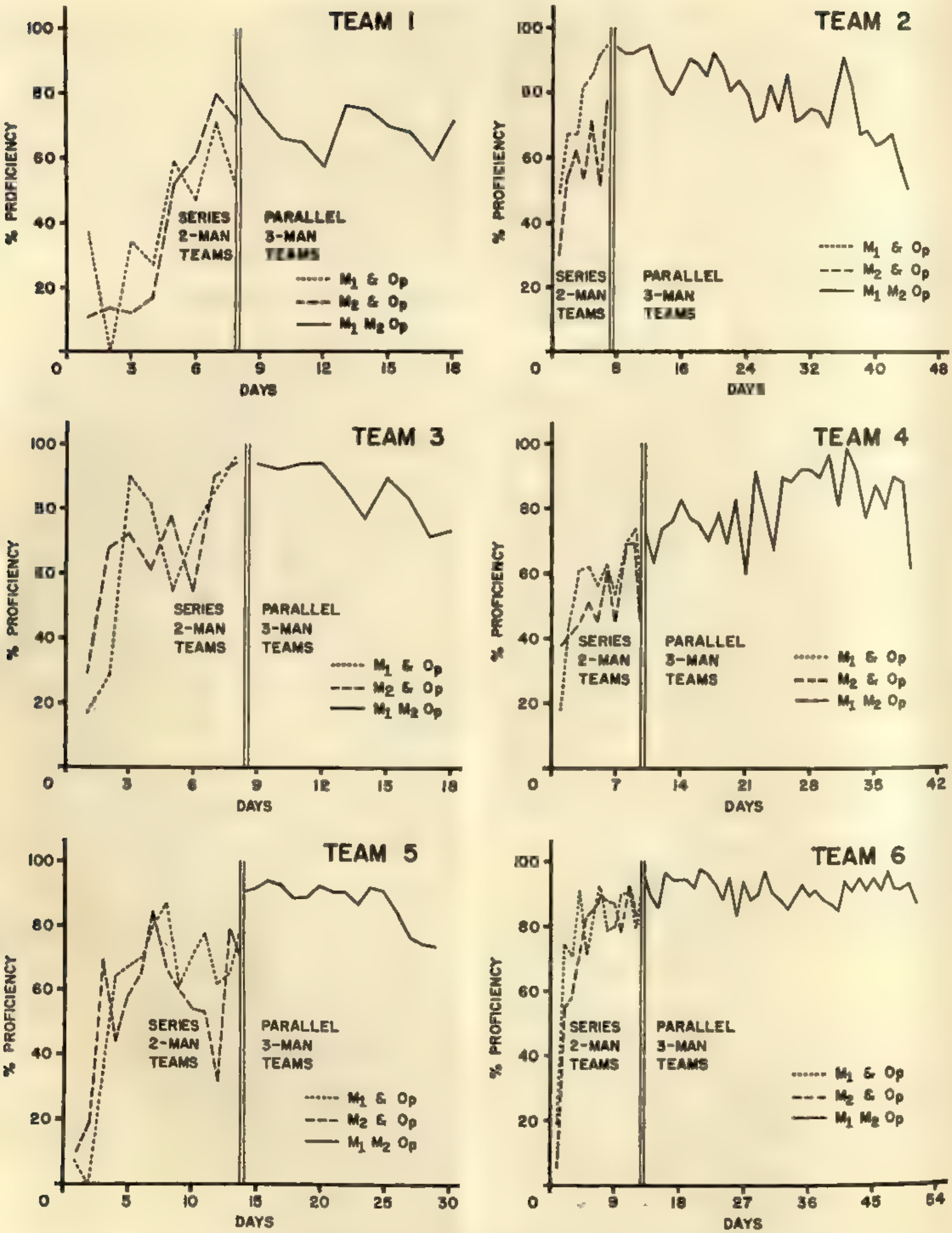

Fig. 5. Percent proficiency for teams in Study II.

training was contingent upon criteria of high and low performance which may not have been representative of each team's overall performance trend, the data in Table 6 describe proficiency during the 12 periods (1 hour) preceding the criterion trials and omit these trials entirely. If the criterion trials had been used as the basis of

comparison, five of the six teams would have shown an increment at the beginning of three-man training and, of course, all teams would have evidenced a decrement at the time this phase was terminated.

It might be argued that evidence of an increment in five of the six teams when the redundant member was added and evidence of a decrement in only four of the six teams when redundancy training was terminated does not support the hypotheses concerning the addition of a redundant member to a team. However, the increment was evident even when the criterion trials were not considered, and the decrement occurred despite additional team practice which, without a redundant member, should have maintained the proficiency of all teams at high levels as a function of continued reinforcement. The question of why all six teams did not decline is discussed below.

Another possible explanation of the decrement is that after 22 to 54 days on the task, the motivation of the subjects was influenced and that boredom occurred leading to performance deterioration. A reasonable refutation of this boredom hypothesis is provided by the data in Table 7. In this table, proficiency data for each monitor is given separately for each quarter of the trials during the sessions of three-man team practice. The column IP describes the proficiency level of each monitor during the period of individual practice which preceded each day's session. As can be seen from this column, individual proficiency was maintained under individual practice conditions

TABLE 6

Study II. Average Proficiency during Noncriterion Trials at the End of Two-man Team Training and under Redundant Conditions

Team	Nonredundant $M_1 + Op$	Nonredundant $M_2 + Op$	Redundant team Initial	Redundant team Final
	12 Trials	12 Trials	12 Trials	12 Trials
1	.55	.69	.82	.89
2	.89	.64	.94	.75
3	.91	.85	.94	.78
4	.71	.59	.63	.81
5	.69	.69	.92	.77
6	.85	.88	.95	.91
Mdn.	.70		.93	.80

throughout the experiment. If boredom did occur, it did not manifest itself during individual practice when continuous reinforcement for correct responses was being supplied. Since there is no evidence that boredom affected individual performance, it is difficult to attribute the team decrement to boredom or other motivational effects.

Results: Molecular Analysis

Detailed results on the performance of the monitors of each of the parallel teams are shown in Table 7. For each monitor, three sets of data are reported for successive quarters of the total number of team trials. The first column indicates the proportion of correct responses made by the monitor during the 10-minute warm-up period of individual practice (IP) preceding each day's team training. The second column indicates the proportion of correct responses made by the monitor during team performance (TP). The third column indicates the ratio of inappropriately reinforcing feedback (IF) to appropriately reinforcing feedback (AF) received during team training. This entry was computed by dividing the number of trials on which the monitor received reinforcement following an incorrect response by the number of trials on which he received a reinforcement following a correct response. The proficiency of the overall team is shown in the last column.

As described in the molar analysis, there was a tendency for overall team proficiency to decline with continued practice by the redundant teams. This is particularly evident in the results for Teams 2, 3, and 5. Team 6, as already reported, evidenced only a slight decline even after 39 days of training. Teams 1 and 4 reached criterion quickly after having maintained or even increased their proficiency. These trends in team performance tend to be reflections of the individual monitor data shown in Table 7. Individual proficiency declined steadily for both monitors in Teams 2, 3, and 5 during team practice (TP). At least one monitor evidenced little change or even an increase in proficiency in each of the remaining teams. Whenever a monitor and his partner both declined in proficiency,

TABLE 7
CHANGES IN INDIVIDUAL AND TEAM PROFICIENCY IN REDUNDANT TEAMS

Team	Days	Q	M_1 IP	M_1 TP	M_1 IF/AF	M_2 IP	M_2 TP	M_2 IF/AF	Overall team
1	11	1	.66	.56	.44	.84	.57	.42	.71
		2	.65	.60	.39	.74	.59	.42	.68
		3	.74	.59	.28	.81	.40	.87	.70
		4	.81	.56	.34	.92	.45	.70	.68
2	37	1	.91	.87	.10	.67	.63	.51	.89
		2	.86	.74	.20	.68	.57	.55	.84
		3	.80	.67	.26	.68	.52	.60	.76
		4	.87	.48	.58	.79	.53	.42	.69
3	10	1	.71	.71	.35	.74	.85	.12	.94
		2	.61	.65	.41	.74	.77	.20	.90
		3	.61	.67	.32	.82	.63	.39	.85
		4	.71	.48	.58	.79	.53	.42	.71
4	30	1	.90	.73	.24	.76	.67	.36	.74
		2	.93	.78	.21	.70	.75	.26	.71
		3	.95	.89	.09	.69	.70	.39	.91
		4	.93	.84	.12	.67	.64	.48	.85
5	16	1	.82	.83	.15	.74	.72	.31	.93
		2	.82	.82	.15	.76	.65	.44	.90
		3	.83	.79	.17	.85	.61	.50	.90
		4	.93	.60	.38	.80	.57	.46	.78
6	39	1	.83	.83	.18	.82	.85	.14	.92
		2	.94	.75	.25	.71	.75	.25	.90
		3	.88	.80	.15	.68	.74	.28	.90
		4	.86	.81	.17	.73	.76	.26	.93

there was a corresponding change in the ratio of inappropriate to appropriate reinforcements (IF/AF). This was the case for Teams 2, 3, and 5. As has been indicated, no decline corresponding to the decrease in team proficiency is evident in individual performance not occurring in the team setting (IP).

The results of Study II have shown that under certain conditions team reinforcement can eventuate in no further increase in correct responding or in a performance decrement both for the team and the individual team members. This outcome is a function of the team arrangement and the probability of correct responses by team members. Specifically, in the three-man redundancy arrangement, reinforcement is contingent upon a correct team response and not necessarily upon correct member responses. This, in turn, permits the strengthening of incorrect responses. In the two-man series arrangement used prior to the parallel team training phase, both members were reinforced whenever the team performed correctly. Under those circumstances, team responses were on a continuous reinforcement schedule but correct individual responses were on an aperiodic reinforcement schedule that depended upon the response probabilities (proficiency levels) of the individual members. With the addition of a redundant member, correct team responses remained on a continuous schedule which, for most teams, initially represented a higher reinforcement ratio than existed in the two-man teams. However, individual reinforcement for the monitors became "confounded" because incorrect responses were aperiodically reinforced. As predicted, whenever the competition offered by the increase in strength of incorrect responses occurred, it tended to result in an eventual performance decrement both for the individual and the team.

This analysis of the effects of adding a redundant member to a team suggests the reasons why Team 6 did not reach the decrement criterion and why Teams 1 and 4 did not demonstrate a stable decline in proficiency. Teams 2, 3, and 5, which most clearly showed the predicted decrement,

were characterized by wider differences in initial monitor proficiencies in the team condition (TP) than Teams 1, 4, and 6. This wide range would facilitate reinforcement for incorrect responses and increase the likelihood of a decrement in team performance as described in the second hypothesis introducing Study II.

The finding that individual proficiency shows differential changes over time depending upon whether performance takes place in an individual or team setting is of interest. With continued practice in the redundancy setting each monitor's proficiency tends to decline while his performance in the individual, 10-minute warm-up sessions tends to remain steady or even increase. It appears that certain differential cues in the two situations, including the conditions of reinforcement, can be discriminated. In the warm-up period, where reinforcement only follows correct responding, proficiency is maintained at a generally high level. However, the team setting where reinforcement also can follow incorrect responding results in more variable and, therefore, less proficient behavior. It is evident that a skill such as "timing" can be performed in different situations with different levels of proficiency.

On the basis of the data in Study II, the following conclusions can be presented:

1. The addition of a redundant member, that is, one whose performance requirement is identical to that of an already existing team member, tends to result in an initial increment in the team's performance. This can be predicted in terms of an increase in the probability of correct performance when a parallel response component is added to the team. The amount of this initial increment depends on the proficiency of the original team. If the proficiency of the original team is high, the addition of a redundant member will have little effect, especially if the redundant member's performance level is low. On the other hand, if the proficiency of the original team is low, the addition of a redundant member will have a decided effect, often even if the redundant member himself has a relatively low proficiency. In general, the relationships between the performance levels of the original team and the subsequent performance levels of a redundant team can be explained in terms of the structure of the team and the individual proficiencies of the members.

2. As training is continued in the redundant situation, performance level is likely to return to or fall below that of the original, nonredundant team. This characteristic of redundant teams is explained by the reinforcing condition which permits aperiodic reinforcement for incorrect member responses. The decrement may develop quite slowly, occurring only after many repetitions of the team task or it may not manifest itself at all depending on how long team performance is observed and on the degree of divergence in the performances of the team members. There were no data in the present study to support the hypothesis that a team might not show a decrement if one member of a divergent team were to perform at a very high level of proficiency. In future studies, the selection of redundant members could be systematically varied in order to control for the amount of difference in their respective proficiencies.

A "commonsense" analysis suggests that one method for increasing team proficiency is to add team members in parallel who would duplicate existing performance requirements. The data from Study II have shown that team proficiency may increase *initially* with the addition of redundant members. However, when this is done, a schedule of reinforcement is introduced in which a member's incorrect responses may be reinforced and an eventual decrement in team proficiency may result. This can occur despite a schedule of continuous reinforcement for correct team responses.

Discussion

Variations in team performance can be described as a function of conditions both external to the group and within it. External conditions refer to events which impinge upon the group from its outside environment; internal conditions refer to the way the group is organized and the manner in

which it functions, for example, its communications, structure, and processing procedures. The research described in this paper provides preliminary evidence for lawful relationships at these two levels: one experiment involved a manipulation of external and one of internal conditions. The first study investigated changes in group performance which occur as a function of the external consequences of group behavior. The second study investigated changes in group performance which occur as the result of an internal change in the manner in which the group is structured.

If the performance of a group is considered without reference to the performance of its members, it is apparent that one group may be better than another at an assigned task or a given group may improve or become worse. The conditions which lead to such changes in group output comprise one level of analysis. A second level of analysis can be thought of as concerned with the effect of the group as an environment in which individual performance occurs. Both the molar and the molecular levels of analysis can contribute to the description of group performance.

Considering only the group and not the responses of the individual members, two generalizations about group performance were identified in this research:

1. The performance of a group is sensitive to, that is, a function of, the consequences of its performance. These consequences were defined as feedback which provides information as to the success or failure of the group's previous actions. This result suggests that group or team practice without appropriate (i.e., differential) feedback will be an insufficient condition to achieve or maintain group proficiency; practice alone may well lead to a decrement in group proficiency as a result of the absence of reinforcement.[7] Even for very high levels of initial team proficiency, some form of differential feedback must be utilized to prevent a deterioration in the quality of group composite responses.

2. While group performance can be related to its consequences, these consequences, in turn, are insensitive to, that is, not necessarily contingent upon, the success or failure of individual member responses. For example, in the first study reported, correct performance by only one participant could not lead to group success even under acquisition conditions. In the second study, incorrect performance by one of the participants often was followed by group success. Furthermore, the group may evidence certain phenomena as the result of the consequences of its performance which are not necessarily consistent with what might be assumed from the study of individual behavior. As was illustrated by the performance decrement of teams in the second study, team performance may deteriorate even when the team is supplied with continuous reinforcement for correct responses.

It is possible to derive these same generalizations about group performance from an analysis of the changes in individual member performance which occur as a function of the reinforcement contingencies experienced by each member. Assuming external conditions remain constant, these contingencies and their influence upon team proficiency can be predicted from a knowledge of the structure of the team and the probability of correct performance by individual members. In this sense, the group constitutes a specifiable environment in which individual performance can be studied.

When an individual performs in a nongroup situation, increments or decrements in his proficiency occur as a result of the reinforcement he receives. The circumstances under which reinforcement is provided are a function of external influences established by the task situation (or by the experimenter in the laboratory) and by the momentary response probability of the individual. In the team setting, the environment external to each member takes on characteristics which reflect the group's structure and the proficiency of other team members. Specifically, in a series team, the schedule of reinforcement for any one member is defined by the probability that all other members are correct whenever he is

[7] Thus, for an athletic team, it is not how well the game is played that counts but whether the game is won.

correct; whether or not he is reinforced following his correct response depends upon the proficiency of the other members of his team. In a parallel team, the predominant condition is the likelihood of reinforcement for an incorrect member response which, again, is determined by the performance of the other team members.

For example, consider a two-man team in which one member has a proficiency of .70 and his partner has proficiency of .40. In a series arrangement, the first member will be reinforced only for 40% of his correct responses; his ratio of reinforcement is determined directly by his partner's proficiency. In a parallel arrangement, his correct responses always would be reinforced. However, he also would receive a reinforcement following 40% of his incorrect responses, again a direct function of his partner's proficiency. As the size of the team increases, or as the level of proficiency of other members decreases, the probability of a reinforcement following the correct response of an individual member of a series team is reduced. In a parallel team, the development of a response by a member which competes with his correct response similarly grows more probable as the size of the team and the proficiency of the other members increases.

To the extent that a team member's environment can be specified in terms of team arrangement (i.e., series or parallel), team size, and member response probability, it is possible to simulate this environment for an individual. This makes it possible to conduct a detailed analysis of the effects of various team-produced learning environments on the performance of individual team members. In turn, the resulting performance of the individual team member will have predictable effects on team proficiency and on the occurrence of team reinforcement. Simulation of the conditions of a group environment can be used to specify the interaction between the influence of a team environment on an individual member and the influence of the performance of that member on the characteristics of his environment. For instance, it is possible to investigate the effect of schedules of reinforcement which are characteristic of particular team arrangements on individual performance and to vary individual performance to determine its effects upon the group environment and upon team proficiency. Some form of simulation, furthermore, can provide opportunities for exerting experimental control over the dynamics (the trial-to-trial changes) of team and team member interactions so as to make the reinforcement contingencies of a team environment for any one trial responsive to the performance of the team member on the immediately preceding trials. The effects of various combinations of team member proficiencies, as these proficiencies change with prior team success, can be controlled experimentally and assessed in terms of the resulting increase or decrease in the probability of a correct individual and team response. Using such an approach, it is conceivable that group phenomena, at least those described in these studies, can be investigated employing "one-member" teams.

In the larger context of social psychology, studies of the kind reported here can provide insights into the differences between a social environment and that traditionally studied in investigations of individual learning. Performance in a social environment ultimately might be described as the consequences of particular reinforcement schedules which are found primarily in multiman systems and which differ from those which are typically investigated when studying individual performance. The methodology of team study reported in this paper might be extrapolated to the complex events characteristic of the tasks and structures of more elaborate social organizations. If the task of social psychology can be defined as the simultaneous investigation of the way in which group membership affects individuals and the way in which individuals influence the groups to which they belong, then this methodology may lead to a better understanding of key aspects of these kinds of behavioral changes.

Summary

The major purpose of the studies described in this monograph was to assess the feasibility of considering a multiman team

as a learning unit which reacts to reinforcement contingencies in much the same way as do individual organisms. Accordingly, a team response should exhibit acquisition and extinction as a function of the properties of the reinforcement situation following each learning trial. Study I was designed to test this general hypothesis by determining the influence of the presence and absence of team reinforcement on the performance of a three-man team in which team reinforcement was contingent upon correct responses from all team members. Study II considered a more complex team structure in which team reinforcement was contingent on correct responses by only some of the team members. This study examined the interactive effects of group feedback upon individual member performance and its subsequent influence on overall team proficiency.

In both studies the units of investigation were three-man teams in which each member was assigned a specific task. The task situation was constructed so that no member received any feedback about the accuracy of his own or any other member's performance until the entire team completed the task. The response required from each team member was the estimation of a short time interval.

In Study I, six teams of three *Ss* each were trained 1½ hours a day for a median of 31.5 days. The first experiment followed an operant conditioning paradigm in which team response learning was investigated in terms of acquisition, extinction, spontaneous recovery, reacquisition, and reextinction. The data obtained from Study I indicated that changes in team performance conformed to predictions made on the basis of knowledge of individual organism behavior when the occurrence of reinforcement is controlled following each response. Using a simple probability model, a predicted level of performance for each team was calculated from individual team member proficiencies. Early in team learning, differences between observed and predicted team proficiency were obtained. This decrement was explained in terms of the change in schedules of reinforcement from the individual to the team setting. With subsequent team practice, observed team performance levels more closely approximated predicted levels.

Study II investigated teams with a parallel structure where the reinforcement contingencies were more complex than for the series teams investigated in Study I. The defining property of the parallel team is the redundancy which exists among team members. Whenever certain team members perform correctly, the performance of other team members is redundant. Common sense often suggests that team errors will be reduced by adding redundant members. However, a redundant member can make an incorrect response and yet be reinforced along with all of the other members of the team following a correct team response. Over a sequence of reinforced trials, these incorrect responses are likely to be strengthened and, on subsequent occasions, such incorrect behavior would have an increased probability of recurring. The primary purpose of Study II was to investigate this property of parallel teams.

The results of Study II showed that the reinforcement contingencies set up by the structure of a parallel team could result in a performance decrement as a function of team arrangement and the probability of correct responses by team members. The data obtained in Study II indicated the following characteristics of parallel team performance. The addition of the redundant member, that is, one whose performance requirement is identical with that of an already existing member, tends to result in an initial increment in team performance. This can be predicted in terms of an increase in the probability of correct performance when a parallel response component is added to the team. As training is continued in the redundant situation, performance level may return to or fall below that of the original nonredundant team. This characteristic of redundant teams is explained by the reinforcing condition which permits aperiodic reinforcement for incorrect member responses. The decrement may develop quite slowly, occurring only after many repetitions of the team task.

The discussion considers changes in team performance as a function of conditions external to and within the team. External conditions refer to events which impinge upon the group from its outside environment and serve as team reinforcers; internal conditions refer to the way the group is organized and the manner in which it functions, for example, its communications, structure, and processing procedures which establish the learning environment for team members. The research in this monograph is seen to provide preliminary evidence for lawful relationships at these two levels.

REFERENCES

GLANZER, M., & GLASER, R. Techniques for the study of group structure and behavior: II. Empirical studies of the effects of structure in small groups. *Psychological Bulletin*, 1961, **58**, 1–27.

KLAUS, D. J., & GLASER, R. *Increasing team proficiency through training: 1. A program of research.* Pittsburgh: American Institutes for Research, 1960.

ROSENBERG, S., & HALL, R. L. The effects of different social feedback conditions upon performance in dyadic teams. *Journal of Abnormal and Social Psychology*, 1958, **57**, 271–277.

(Received July 7, 1965)

Vol. 80, No. 14 Whole No. 622, 1966

Psychological Monographs: General and Applied

EXPERIMENTAL ANALYSIS OF RESPONSE SLOPE AND LATENCY AS CRITERIA FOR CHARACTERIZING VOLUNTARY AND NONVOLUNTARY RESPONSES IN EYEBLINK CONDITIONING[1]

KENNETH P. GOODRICH

Macalester College

Previous work on the identification of instrumental or voluntary (V) responses in eyeblink conditioning led to the use of a response latency criterion in some experiments and a response slope criterion in others. The present study (a) examines the relation between response latency and slope and (b) seeks by instructing some Ss to blink and others not to blink to develop rational criteria for the identification of V responses. Latency and slope were clearly not equivalent bases for this identification. Moreover, analyses of the data of Ss who received the special instructions showed that the conventional latency and slope criteria both had serious deficiencies. New criteria developed from these data were more successful but still of debatable value. The implications of these findings for the significance of eyeblink conditioning research were discussed.

When defined in terms of the procedures which experimenters follow, classical and instrumental conditioning experiments are distinctly different. In the procedure which defines instrumental conditioning, some object or event is made an outcome for some response of the subject (*S*). This outcome, which is controlled by the experimenter, comes to "control" the behavior upon which it is contingent. In the procedure which defines classical conditioning, no such contingent outcomes, or instrumental contingencies, are specified. The conditioned stimulus (CS) and the unconditioned stimulus (UCS) are presented to *S* regardless of what *S* happens to be doing.

Psychologists have not generally been content with this procedural distinction, however unambiguous it may be. They have wanted to know whether different learning processes correspond to the different procedures. Both negative and affirmative answers have been given to this query, and a variety of mechanisms have been proposed to make either answer consistent with the presence of contingent outcomes in instrumental conditioning experiments and their absence in classical conditioning experiments.

We must note carefully at this point that the absence of contingent outcomes in classical conditioning experiments is an absence which occurs in the "rules" followed by the experimenter. Such an absence of contingent outcomes does not necessarily mean that effective temporal contiguities between responses and controlling outcomes are absent in a particular classical conditioning experiment and not playing a major role in determining the behavior which occurs in the situation. Hull (1943, pp. 74–79), for example, argued that since an unconditioned response (UCR) occurred in time after a CS and was accompanied by the UCS, his theory of instrumental learning could be applied to the classical conditioning experiment if the UCS were properly construed as a drive reducer. That is, an outcome (the presence of a positive UCS or removal of a negative UCS) follows the response and thus helps to attach the response to the CS, in spite of the fact that the experimenter presumably does not explicitly arrange an

[1] This work was partially supported by National Science Foundation Grant GB-203 to the University of Pennsylvania and by National Institutes of Health Grants MH-04528 and FR-00167 to the University of Wisconsin. Alice Isen, Nancy Miller, and Susan Shuben ran the *S*s and did preliminary data tabulation. F. Robert Brush, Joseph Markowitz, and A. Martin Wall provided valuable assistance and criticism at several points during the investigation.

instrumental contingency. Hull, then, did not believe there was a classical conditioning *process*. The procedural difference was not one that made any fundamental difference. Without outcomes, responses would not be learned.

Hull's viewpoint has not been widely adopted, perhaps in part because of the experiments by Mowrer and his associates (Mowrer & Aiken, 1954; Mowrer & Solomon, 1954) which have generally been interpreted to cast doubt on Hull's interpretation of classical conditioning. Thus, it is still a plausible working hypothesis that a classical conditioning *process* exists which does not require outcomes for the learned response. From this point of view, the results of a classical conditioning *experiment*, as defined above, do not necessarily provide a clear picture of the classical conditioning *process*. To provide such a picture, it must be possible to show that there are no inadvertent, confounded, or "superstitious" temporal contiguities between response and outcome in any given classical conditioning experiment.

One of the most widely employed classical conditioning situations is eyeblink conditioning, as shown by even a hasty survey of the literature over the last 25 yr. and by the fact that many of the principles of conditioning discussed by Kimble (1960) gain a large measure of their support from such experiments. To the extent that one wants to regard these experiments as reflecting the *process* of classical conditioning (in the sense discussed above), one must argue that the temporal contiguities discussed above cannot reasonably be said to contribute to the essential features of the results.

It would appear that of the several classical conditioning situations most frequently encountered, eyeblink conditioning is the one for which it is most difficult to dismiss the possible role of confounded contiguities between response and outcome. To one approaching eyeblink conditioning for the first time, whether a psychologist or a college sophomore serving as an *S* in such an experiment, the conclusion often seems inescapable that one is dealing with instrumental avoidance conditioning: by giving an anticipatory blink to the CS *S* is mitigating the unpleasant effects of a puff of air in the eye. Certainly there is no doubt that the eyeblink can be an instrumental act. We will most likely observe an instrumental blink if we walk up to someone and, by means of a threat, make avoidance of a poke in the nose contingent on a blink. The question, then, is not whether the eyeblink *can* be brought under the control of its outcomes but rather whether in the classical conditioning situation it *is* under such control.

Voluntary and Conditioned Responses

Given the possibility that the avoidance contingency does play some role in eyeblink conditioning, what course of action is indicated for one who wishes to study the classical conditioning process as discussed here? One line of work which has come to grips with this problem has attempted to distinguish *voluntary* (V) from *conditioned* (C) eyeblinks and to find ways of removing from conditioning data the V responses. It should be noted at this point that the assumption is made that "voluntary" in this context means much the same as "instrumental" in our earlier discussion. Moreover, it is assumed that the great majority of the voluntary or instrumental responses which one would encounter in an eyeblink conditioning experiment would arise through the avoidance possibility. It is likely, of course, that V responses will arise for other reasons as well, such as misunderstanding of instructions or motives to "cross-up" the experimenter. The controlling outcomes for such responses are idiosyncratic and not directly a function of the experimental procedures. These "unconfounded" outcomes are contrasted with the "confounded" outcomes produced by avoidance: the possible effects of an anticipatory response on the aversiveness of the UCS which follows it.

The work reported in the present paper follows directly from previous work by Spence and Ross (1959) and Hartman and Ross (1961). These two papers, in turn, were based on earlier observations by Spence and Taylor (1951). Spence and Taylor proposed a procedure for eliminating V

responses from the data obtained in eyeblink experiments. On the basis of observations which were not presented explicitly, they concluded that *Ss* who admitted to being aware of the purpose of the experiment typically gave responses with a characteristic form and with latencies shorter than 300 msec. Thus Spence and Taylor proposed that in a conditioning experiment, an *S* whose median latency of anticipatory responses was less than 300 msec. be discarded.

Before continuing this historical introduction, we may pause to consider briefly the reason for discarding *Ss* rather than responses in order to rid data of voluntary influences. The answer appears to lie in the reasonable presumption that the occurrence of a V response on a given trial will preclude, or at least greatly attenuate the probability of, the occurrence of a C response on that trial. Therefore an estimate of the probability of a C response obtained with data from which V responses have been removed would be spuriously low. Because learning curves expressed in terms of estimated probabilities of responses have generally been the main object of concern in eyeblink conditioning experiments, *Ss*, rather than responses, have been eliminated. The reader should remember, however, that the validity of such a procedure has ultimately been evaluated in terms of how successfully it manages to eliminate V responses.

For several years following the work of Spence and Taylor, the procedures recommended by these authors were adopted in many eyeblink conditioning experiments. In 1959, Spence and Ross published an article in which data were presented in support of the latency criterion for identifying V responses. Two independent observers examined each response of each *S* in a conditioning experiment and on the basis of the form of the eyelid closure judged the response to be voluntary, conditioned, or not scorable. The criteria employed by the judges (discussed in greater detail below) presumably were based on responses of an unknown number of *Ss* who had been instructed to blink; these data were not presented. However, when frequency distributions were plotted for the latencies of the responses in the several judged categories, the latency procedure was shown to be a good one. That is, the majority of responses judged to be voluntary were eliminated by discarding *Ss* whose median latencies were less than 300 msec. It was argued that the latency distinction was justified insofar as it resulted in the elimination of responses which were judged to be voluntary (and the simultaneous retention of most of the responses judged to be nonvoluntary).

Although various investigators continued to use the latency discard criterion, no information was available concerning whether this criterion worked in the face of various kinds of experimental variations. In 1961, Hartman and Ross reported that for one such variation the latency criterion was quite inadequate. The latency procedure had been developed at the University of Iowa where a ready signal was always employed in eyeblink conditioning experiments. Such experiments at the University of Wisconsin typically did not employ a ready signal (e.g., Grant & Schipper, 1952). Following his move to the University of Wisconsin, Ross collaborated with Hartman in seeking the answer to the question of whether the latency criterion worked as well without the ready signal as it did with it. Their answer was that it did not. The results of their experiment showed that V responses without a ready signal occurred with longer and more variable latencies than the V responses reported in the Iowa studies. Hartman and Ross argued that this was entirely reasonable since, as in a reaction-time experiment, a ready signal would permit *S* to be "set" and therefore to make his V responses faster. Without a ready signal, responses would vary in latency and more of them would occur with latencies longer than 300 msec.

Hartman and Ross proposed an alternative discard procedure. Noting that the judgments of response types were based on the shape or form of the response, they proposed that an objective measure of the rate at which the eye closed be used in differentiating V from C responses. Specifi-

cally, they proposed that any S be discarded whose median relative response slope was greater than 40%. The relative slope of any anticipatory response was defined as its maximum slope divided by the mean maximum slope of that S's UCRs on the first five trials. Hartman and Ross went on to show that if Ss were eliminated by this slope criterion the data were purified of judged V responses to about the same extent as had been true in the Spence and Ross experiment with the latency criterion.

In summary, work to date has made available two objective criteria for eliminating V responses from eyeblink conditioning data. One, used in experiments with a ready signal, employs the latency of the response. The other, used in experiments without a ready signal, employs the slope of the response. Several important questions have not been answered. What is the relation between the slope and the latency of the response? The presence or absence of a ready signal apparently affects response latency. Does it also affect response slope? Are slope and latency equivalent bases of classification when a ready signal is employed?

The Present Paper

In the first part of the present paper a detailed examination is made of standard conditioning data, obtained with a ready signal, in terms of both the conventional latency and slope criteria for eliminating V responses. This analysis indicates certain discrepancies between the results of applying the latency and slope criteria. These discrepancies lead to an attempt to isolate experimentally pure cases of V and C responses and Ss in order to derive meaningful criteria. Following a reanalysis of the conditioning data in terms of some new procedures, the current status of eyeblink conditioning is evaluated in light of the available methodology for identifying voluntary processes.

The results of two experiments are reported. The second experiment in large measure constitutes a replication of the first. Included with the data from the first experiment are data from a study conducted for quite a different purpose by Harold Fishbein as part of a doctoral dissertation (Fishbein, 1963). The analyses of Fishbein's experiment which are reported here were not a part of his thesis but were performed later on data generously supplied by Fishbein. The results of the two experiments will be presented separately but concurrently. In this way certain differences will be seen which provide some information about the effects of sampling and minor procedural changes, as well as some important invariances.

METHOD

Apparatus

The S was seated in a dental-type chair within a sound-insulating cubicle. Air conditioning fans produced an ambient noise level in S's cubicle of 60 db. (re .0002 dynes/cm^2). The E was located in another room with the apparatus for controlling stimuli and recording responses.

Each S received 60 trials which occurred with intertrial intervals of 15, 20, and 25 sec. in an irregular sequence. A set of electronic interval timers controlled the stimulus presentations on each trial. The first event on any trial was a ready signal which consisted of a 400-msec. burst of white noise (80 db.) from a speaker over S's head. Following a 2-, 3-, or 4-sec. interval, a visual CS was presented for 550 msec. on a display panel in front of S. The CS was produced by lighting a neon lamp (GE NE40) behind a 19.5 × 19.5 in. white translucent screen, causing a change in luminance of a 2.5-in. diameter circular area on the screen from a background level of 0.03 mL. to a stimulus level of 1.73 mL. In Fishbein's experiment and in Experiment 1, the CS appeared as an illuminated spot on a plain field. In Fishbein's experiment the spot occurred in the center of the field; in Experiment 1 it appeared either at the left or at the right in a balanced sequence (cf. Goodrich, 1964b). In Experiment 2, the CS appeared as the illumination of a circular portion of screen defined by a 2.5-in diameter cutout in the center of a black ground which masked the remainder of the screen.[2]

The UCS was a 50-msec. puff of air which arrived at S's eye 500 msec. after the onset of the CS. To encourage voluntary responding, its intensity was set at the rather high level of 5 psi, measured at the air tank some 15 ft. from S's eye.[3]

[2] Results reported elsewhere by the author (Goodrich, 1964b) make it unlikely that differences in the results among the subexperiments here were the result of differences in the CS display.

[3] That more voluntary responding occurs with stronger air intensities was assumed by Spence and Ross (1959) and is generally consistent with the findings of Gormezano and Moore (1962).

The air was conducted through 3/16-in. copper and glass tubing and was released with a solenoid valve in the line some 8 ft. from S's eye. The nozzle was positioned so as to direct the air at the corner of S's right cornea from a position just clear of the lashes.

Responses were recorded with a system which employed a microtorque potentiometer coupled to the lid of S's right eye with a light thread, a plastic "false eyelash," and a strip of adhesive tape. Voltage changes produced by resistance changes in the potentiometer were amplified and recorded on a Brush oscillograph in the manner described by Goodrich, Markowitz, and Norman (1964).

Before running each S the relation between stimulus events and the corresponding polygraph tracings was checked. Also checked was the calibration of the interval between light onset and arrival of the air puff at S's eye.

The Experimental Conditions

The several experimental treatment conditions were differentiated largely by the instructions to S concerning the task. These conditions were as follows:

Conditioning. The instructions for Ss in this condition were essentially those used in the Iowa laboratory. Briefly, they informed S that E was concerned with studying "attention and readiness," that this was achieved by observing the reaction of S's eyes to stimulation, that a light stimulus would come on and a puff of air would be delivered to the eye, that a ready signal would occur and was to be followed immediately by a voluntary blink, and that S was not to try to control the responses of his eye to stimulation but let the responses of his eye "take care of themselves." Finally (as if this were possible), S was asked to relax and think about other things.

Instructed-inhibit. The Ss in this condition received essentially the same instructions as just described with the very important difference that they were told that their task was *not* to blink while the light (CS) was on. They were told that following each trial they would be informed, by the onset of one of two lights beneath the stimulus display panel, whether they had succeeded in not blinking. In addition, they received a reward for each success ($.01 in Fishbein's experiment, $.05 in Experiment 2) and a punishment for each failure (loss of $.01 in Fishbein's experiment and of $.05 in Experiment 2). The Ss in Experiment 2 started with a credit of $2.00; those in Fishbein's experiment started with $1.00.

Instructed-blink. The Ss in this condition also received the basic instructions, with two important differences. First, these Ss were instructed to blink "while the light is on." Second, in one subcondition of the Instructed-blink condition the UCS was not employed. For these Ss, the nozzle was adjusted to the eye but no puff was ever delivered during the experiment and the puff was not mentioned in the instructions. The Ss in the remaining subcondition did receive the puff and the appropriate puff instructions.

General Procedures

The S was taken into the sound cubicle and the headpiece was adjusted with no instructions and a minimum of informal conversation. Although E made no attempt to be austere or to frighten S, the experience probably was unsettling for most Ss, as Spence (1964) has suggested. The instructions were then read and S was requested to explain to E what he was to do and what would happen. The E reread or paraphrased portions of the instructions if S appeared not to understand. During the experiment proper, no questions concerning the purpose of the experiment were answered. The E entered S's room only on the rare occasions when the polygraph record indicated that the linkage between S's lid and the potentiometer was out of adjustment. At the end of the experiment, E attempted to get S to talk briefly about what S had been trying to do and what he thought the experiment was about. Beyond reporting that a wide variety of answers was given to this questioning, ranging from apparent failure to note anything during the experiment to a rather precise analysis using the terminology of conditioning, no further use will be made here of this rather unsystematic and incompletely recorded information.[4]

Analysis of Data

The raw data consisted of polygraph tracings. Several sample recordings are reproduced in Figure 1. Each of the 12 panels shows a single trial, beginning just prior to CS onset and ending about 100 msec. following UCS onset. An anticipatory response was defined as the occurrence of a downward deflection of the recording pen at least 1 mm. below the eye-open baseline in the 500-msec. interval following the onset of the CS. Since the recording gain was calibrated to yield one-to-one recording,

[4] The writer has informally compared two ways of questioning S at the termination of an experimental session. In the first, S is asked in the usual way what he thought the experiment was about and what he was trying to do. In the second, S is told that his data will be of no value unless he knows what the experiment is about. The second mode of questioning seems to result in a very high incidence of Ss who describe rather accurately the classical conditioning nature of the experiment. Although these observations were not as systematic as one could wish, and although S may figure out the experiment only after it is over under the second method of questioning, the writer believes that Ss normally know far more than they admit to knowing.

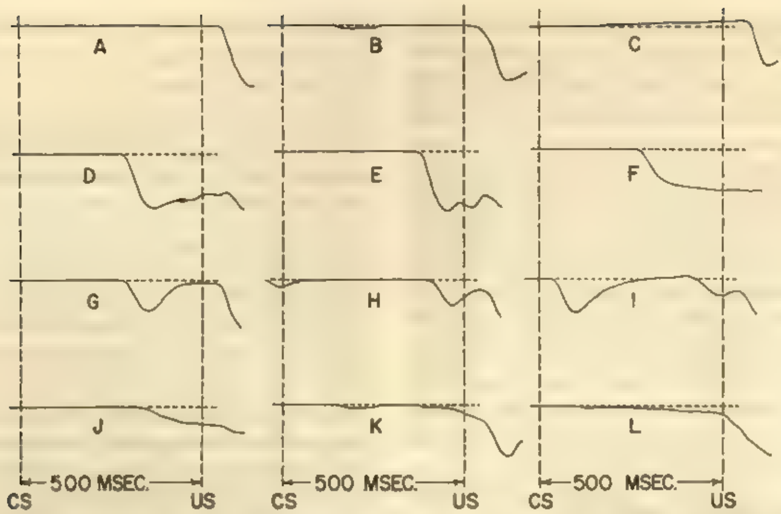

FIG. 1. Sample records of a variety of response patterns. Each record is a polygraph tracing from one trial.

the response criterion corresponded to a lid movement of 1 mm. Records B and K in Figure 1 show anticipatory responses with amplitudes of approximately 1 mm.

Two scores were derived from the polygraph tracings: (*a*) the latency of each anticipatory response was measured in terms of the number of millimeters separating CS onset from initiation of pen movement on the record (8.1 msec. per millimeter); (*b*) a measure of response slope was also recorded. In pilot work, the determination of maximum slope with a protractor and a table of tangents proved to be an extremely tedious task. (Hartman and Ross had used an electronic differentiator which provided an oscillograph tracing the amplitude of which was proportional to slope.) As an alternative, a measure was defined which could be obtained at the same time as the measure of latency with a transparent scoring template placed over the polygraph record. This measure was the distance the eye had closed 50 msec. following the initiation of the blink.

In the case of brief responses which had clearly reached a maximum excursion prior to 50 msec. after their initiation, a linear extrapolation out to 50 msec. was made using as references the point of initiation and the point at which the maximum excursion had been achieved. In Figure 1, responses such as those in Records B and K, if at least 1 mm. in amplitude, would have been treated in the manner just described. For other shallow responses, such as the one shown in Record L, no extrapolation was necessary because the pen was still moving downward 50 msec. following response initiation.

Following Hartman and Ross, the excursion measure for each anticipatory response was expressed as a fraction of the measure of that *S*'s UCR in order to compensate for variability among *S*s in the rate of typical eye closures. The measures of the UCR were obtained on two extra trials, Trials 61 and 62, with an air puff but no CS or additional instructions. If one or both of these UCRs was confounded with an anticipatory blink, as in the case of many of the sample records in Figure 1, one or both of the UCRs from Trials 1 and 2 were used. The measure of a complete UCR, an example of which is shown in Record A of Figure 1, was generally in the range of 12–15 mm. The measure for anticipatory responses ranged from about 1.5 mm. (responses like those shown in Record K) to values similar to those of UCRs (responses like those in Records D and E).

A scatter-plot is presented in Figure 2 which relates relative excursion values (as defined above) on the ordinate to the rela-

tive tangents employed by Hartman and Ross on the abscissa for the responses of 10 *S*s from an independent experiment conducted by the writer. The correlation is clearly high and the proportionality constant not too different from unity. As a further check on the comparability of the present slope procedures with those of Hartman and Ross (1961), a post hoc comparison was made, using the data from Conditioning *S*s in Experiment 2, of the UCR excursion measures based on Trials 61 and 62 with excursion measures based on Trials 1–5. The reason for this comparison is that Hartman and Ross had used Trials 1–5 to obtain UCR slope values whereas Trials 61 and 62 were used in the present study. The mean excursion measures were 12.8 mm. and 12.6 mm. for Trials 1–5 and Trials 61–62, respectively, and the standard deviation of the 29 differences was 1.4 mm. It may be concluded that biases were probably not introduced by using Trials 61 and 62 instead of Trials 1–5.

The term "slope" in this paper will refer to the relative excursion measure discussed above, and the comparability of this to the measure devised by Hartman and Ross will be assumed.

Subjects

All *S*s were undergraduates at the University of Pennsylvania. The data of 41 *S*s were not used. In Experiment 1, 28 *S*s were discarded because of procedural errors or apparatus failures, 3 because of failure to give at least one anticipatory response, and 1 because of an excessive rate of random blinking. In Experiment 2, nine *S*s were discarded, all because of procedural errors or apparatus failures. Fishbein (1963) reported no discarding of *S*s.

The remaining *S*s were distributed among the experimental conditions as follows.

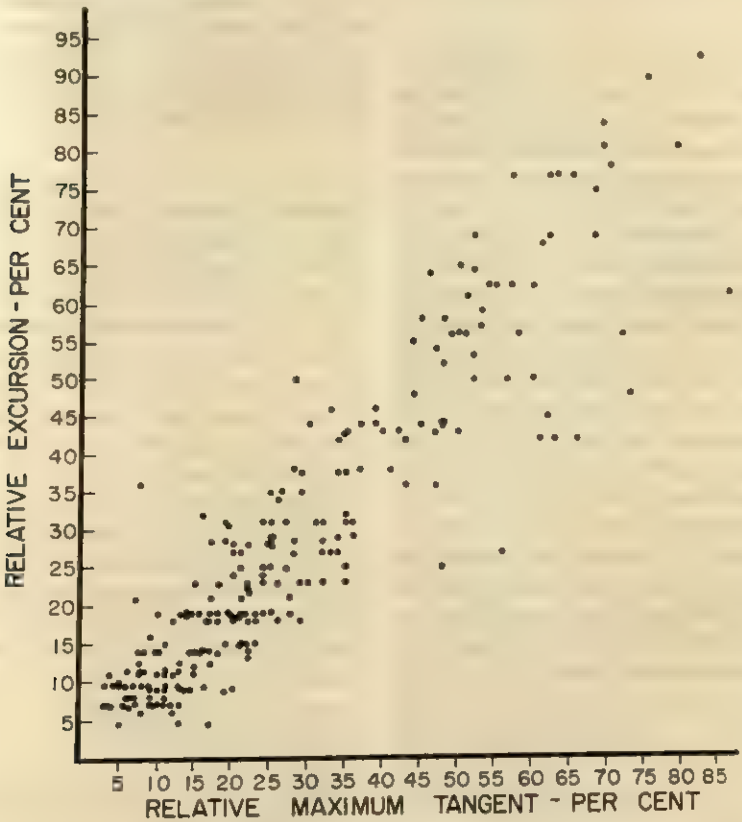

FIG. 2. Scatter plot pooled over 10 *S*s in a separate conditioning experiment showing the covariation of the relative maximum tangent and the relative excursion of anticipatory responses.

Thirty men served in Fishbein's Instructed-inhibit group. In Experiment 1, 42 men served under conditioning instructions, and 78 men served under instructions to blink, 36 with the UCS and 42 without the UCS. In Experiment 2, 52 *Ss* (34 men and 18 women) served under instructions not to blink, 18 *Ss* (11 men and 7 women) served under instructions to blink (without the UCS), and 29 *Ss* (20 men and 9 women) served under conditioning instructions.

The *Ss* in Experiment 1 "volunteered" to serve without pay as a course requirement. The Instructed-blink and Conditioning *Ss* in Experiment 2 volunteered to serve for a fee of $2.00. The *Ss* who served under instructions not to blink were volunteers who received a variable sum of money which depended on how well they succeeded in not blinking (as described above).

RESULTS

Slope and Latency Analyses of Conditioning Data

Most of the analyses in the present report are based on frequency distributions of response latencies and slopes. Many of these distributions were compiled by combining data over both *Ss* and trials, a procedure generally followed in earlier work. Clearly, errors from averaging may be introduced by this procedure. Nonetheless, for individual *Ss* and small blocks of trials there were insufficient numbers of responses to make reliable inferences about the shapes of frequency distributions. In a later part of this report there will be occasion to comment on the relation in the present data between individual and grouped distributions.

Distributions for Trials 1–30 and Trials 31–60. Figure 3 presents the frequency distributions of response latency for *Ss* run under conditioning instructions in the two experiments. To show how these distributions changed with trials, the distributions were plotted separately for Trials 1–30 and Trials 31–60. Between the first half and the second half of the training sessions there was a decrease in the frequency of responses with latency less than about 150 msec. and an increase in the frequency of responses with latency greater than about 200 msec. During the second half of the session the principal mode of the latency distributions occurred somewhere after 400 msec., and there was a clear skewness to the left. The secondary mode, occurring in the region of around 120 msec., presumably resulted from responses occurring as unconditioned reactions to the CS. This mode had largely disappeared by the second half of the training session.

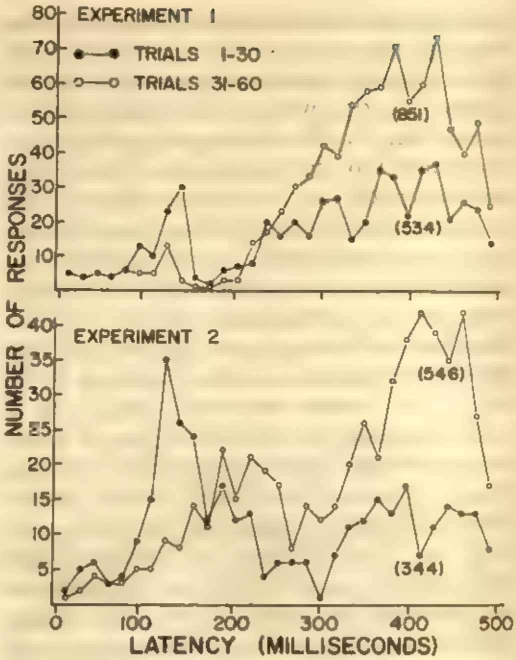

FIG. 3. Distributions of response latencies for the Conditioning *Ss* in Experiment 1 ($N = 42$) and Experiment 2 ($N = 29$) plotted separately for Trials 1–30 and Trials 31–60. Numbers in parentheses in this and succeeding figures are numbers of responses in distributions.

The properties just presented describe the results of both Experiment 1 and Experiment 2, and are similar to those reported by other investigators. In one respect the results of the two experiments seem to differ. In Experiment 1 there was a single major mode in the region after about 150 msec. In Experiment 2, on the other hand, there apparently were two modes in this region, one occurring in the vicinity of 250 msec. and the other in the region of the corresponding mode in Experiment 1. The latency value previously used to differentiate

V from C responses was 300 msec., a point which fell roughly between the two modes in the results of Experiment 2 but had no apparent significance with respect to the distribution in Experiment 1.

In Figure 4 are presented the slope distributions for the two experiments plotted separately for Trials 1–30 and Trials 31–60. Here may be noted a major mode in both experiments in the region of 15% relative slope, and a marked skewness to the right. The shape of these distributions did not appear to change systematically with the addition of more trials. The slope value used to differentiate V from C responses in previous work was 40%. Only in Experiment 2 was there any suggestion of a possible bimodality about the 40% point.

Distributions of V and C responses. The application of the conventional latency and slope criteria for differentiating V and C responses is shown in Figures 5 and 6. To minimize the effect of the unconditioned responses to the CS, the data in these figures are based on only the last half of the training session.[5] Figure 5 presents the frequency distributions of response latencies for responses identified as voluntary and conditioned according to the 40% slope criterion. Several features of the data in Figure 5 are consistent with the previous work by Spence and Ross (1959) and Hartman and Ross (1961). In particular, we note the occurrence of more C-slope then V-slope responses in the portion of the distribution beyond 300 msec., and more V-slope than C-slope responses prior to 300 msec. The latter effect is most marked in the data from Experiment 2. These distributions are not

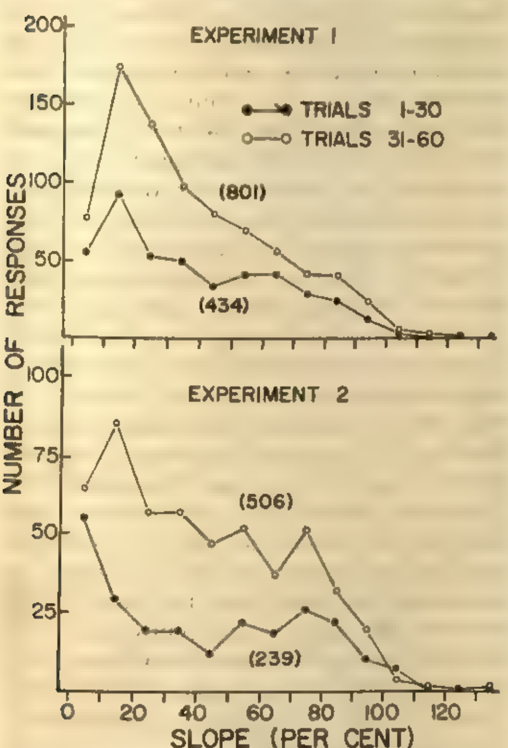

FIG. 4. Distributions of response slopes for the Conditioning *Ss* in Experiment 1 ($N = 42$) and Experiment 2 ($N = 29$) plotted separately for Trials 1–30 and 31–60. Data include only responses with latencies greater than 150 msec.

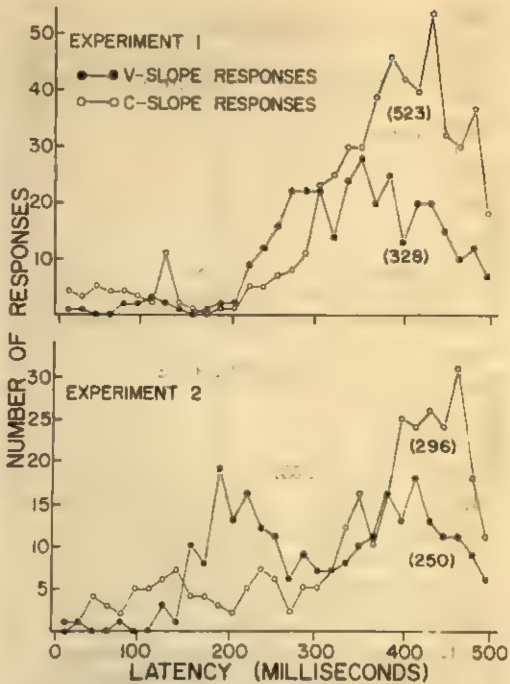

FIG. 5. Distributions of response latencies for V-slope and C-slope responses made by the Conditioning *Ss* in Experiment 1 ($N = 42$) and Experiment 2 ($N = 29$). Data were obtained from Trials 31–60.

[5] We may note in passing a confirmation of other work (Goodrich, 1964a) showing that the original responses to the CS are predominately shallow in slope. Thus they resemble in this respect the C response rather than the V response or the UCR.

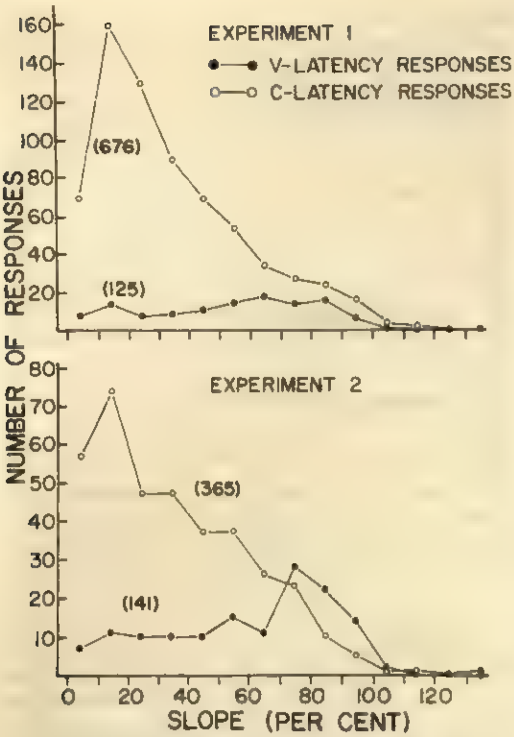

FIG. 6. Distributions of response slopes for V-latency and C-latency responses made by the Conditioning *S*s in Experiment 1 ($N = 42$) and Experiment 2 ($N = 29$). Data were obtained from Trials 31–60 and include only responses with latencies greater than 150 msec.

precisely comparable to the latency distributions presented by Spence and Ross and by Hartman and Ross, since in the work of these investigators responses were differentiated on the basis of judged categories whereas in the present study they were differentiated on the basis of measured response slope. Hartman and Ross, however, demonstrated good agreement between judged form and measured slope.

The presence of a mode for V-slope responses in the vicinity of 200 msec. in Figure 5 is consistent with the latency distributions reported by Spence and Ross in which responses judged to be voluntary had a mode between 200 and 300 msec. Nonetheless, the distributions shown in Figure 5 contrast in an important respect with the data of Spence and Ross. A sizable proportion of the V-slope responses occurred beyond 300 msec. in the present experiments (69% and 55%), whereas a much smaller proportion occurred in this region in the data of Spence and Ross (less than 17%). In this respect, the present results resemble more closely the results reported by Hartman and Ross than those reported by Spence and Ross, in spite of the fact that the latter employed a ready signal as in the present study. Several other sets of data obtained by the writer have consistently shown a sizable proportion of V-slope responses in the area beyond 300 msec. Although it has recently been confirmed that a ready signal produces more V-slope responses with short latencies than does no ready signal (Goodrich, 1964a), average latency apparently is not invariant with other, unknown, variations among experiments.

Frequency distributions of response slopes are presented in Figure 6 for responses identified as voluntary and conditioned on the basis of the 300 msec. latency criterion. The great majority of responses in both experiments were identified as conditioned rather than voluntary according to this criterion. These C-latency responses occurred with greatest frequency in the region of rather shallow slopes and decreased in frequency to produce a marked skewness toward the right. In contrast, the smaller number of responses identified as voluntary had a slope distribution of quite a different shape. If any mode may be identified, it clearly lies toward the steep end of the slope continuum rather than toward the shallow end. This is particularly clear in the case of Experiment 2 where a mode occurred around 75% slope. Examination of these data showed that the responses which produced the mode at 75% in the slope distribution were generally the same responses which produced the mode at 200 msec. in the latency distribution. For unknown reasons, this phenomenon was absent in Experiment 1; only Experiment 2 produced a sizable group of short-latency, steep-slope responses.

Relation between slope and latency criteria. It is clear from the preceding two figures that although in several respects our distributions of slope and latency are simi-

lar to those previously reported, there are a disturbing number of instances in which responses meet only one of the two conventional criteria and not both. These findings are summarized in Figure 7, which presents the relative number of responses in the four possible categories formed by jointly classifying responses according to both latency and slope. Approximately 65% of the responses were categorized alike by the two criteria; 35% of the responses thus met one but not both of the criteria. The more numerous type of inconsistent classification was that involving V slope and C latency. This finding reflects the same fact referred to above in our examination of the latency and slope distributions: the presence of a large number of responses which had slopes greater than 40% and latencies longer than 300 msec.

When the conventional criterion for identifying voluntary *S*s (those with at least 50% V responses) was applied, the results resembled those in Figure 7. Several instances were found in which *S*s met either the slope or latency criterion, but not both. Approximately 75% of the *S*s were categorized alike by the two criteria; 25% were inconsistently classified. In all cases, the latter *S*s were voluntary by the slope criterion but not by the latency criterion.

Effect of discarding V Ss. The data presented in Figure 7 were based on the entire set of responses in each of the two experiments. We turn now to an examination of the effects upon the relative frequency of the four kinds of responses as *S*s were eliminated who failed to satisfy various criteria for inclusion in the sample. Figures 8 and 9 summarize the relevant data. Figure 8 represents the effects of discarding *S*s on the basis of response latency. In this case an *S* would be discarded from the experiment if a certain proportion of his response latencies fell in the region below 300 msec. The right end of each abscissa in Figure 8 corresponds to the stiffest of all criteria. At this point all *S*s would be discarded who had 0% or more V-latency responses. Moving to the left on the abscissa the discard criteria become successively more lenient, until at the far left end of the scale an *S*

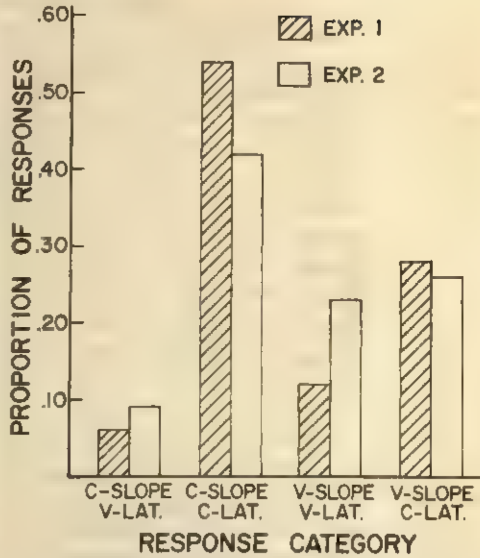

FIG. 7. Proportions of responses falling in the four categories formed by jointly classifying responses by the conventional slope and latency criteria. Data were obtained from the Conditioning *S*s in Experiment 1 ($N = 42$) and Experiment 2 ($N = 29$) over Trials 1–60. Only responses with latencies greater than 150 msec. were included, 1235 in Experiment 1 and 745 in Experiment 2.

would be discarded only if 100% of his responses were voluntary by the latency criterion. The top panels in Figure 8 reflect the gross effects of applying the various discard criteria. The filled circles show the proportion of total responses which would remain after application of the criteria on the abscissa. The open circles show the proportion of *S*s which would remain.

The lower panels in Figure 8 show the relative frequency of the four kinds of responses in the samples which remain after applying each discard criterion.[6] At the far left, where almost no *S*s are discarded, we see proportions which are nearly identical to those previously presented in Figure 7. Moving to the right, it is apparent that as the discard criterion becomes more stringent

[6] Spence and Ross (1959) and Hartman and Ross (1961), in analogous plots, presented the relative frequencies of V and C responses in the data which were *eliminated* by the various discard criteria. It would seem more important, however, to be concerned with the composition of the data which *remain*.

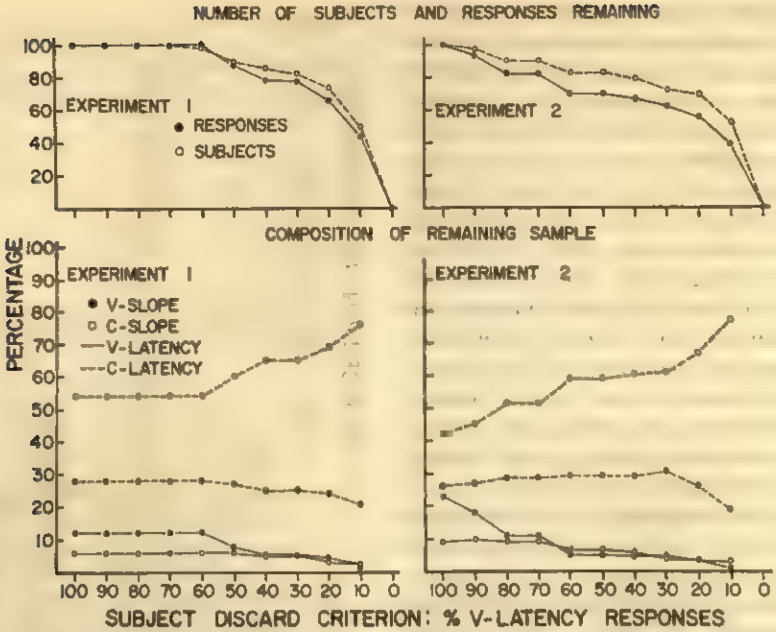

FIG. 8. Effects of discarding Conditioning *Ss* with the conventional latency procedure upon the amount and composition of data remaining. Data were obtained over Trials 1–60 and include only responses with latencies greater than 150 msec., 1235 in Experiment 1 and 745 in Experiment 2.

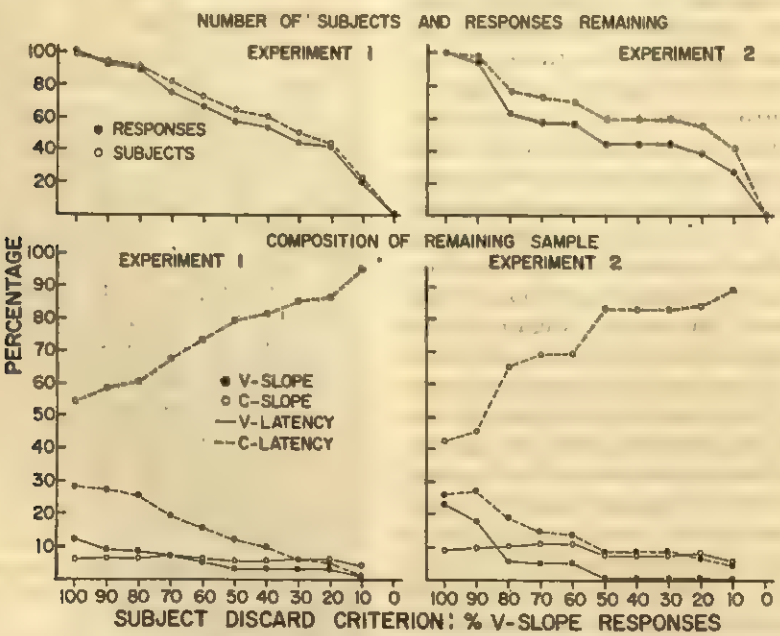

FIG. 9. Effects of discarding Conditioning *Ss* with the conventional slope procedure upon the amount and composition of data remaining. Data were obtained over Trials 1–60 and include only responses with latencies greater than 150 msec., 1235 in Experiment 1 and 745 in Experiment 2.

in both experiments the relative frequency of responses which are classified as conditioned by both slope and latency increases, and the relative frequency of responses which are classified as voluntary by both slope and latency decreases. At the 50% discard criterion recommended by Spence and Ross and by Hartman and Ross, the relative frequency of C-slope, C-latency responses has been increased to approximately 60%, and the relative frequency of V-slope, V-latency responses has been reduced to approximately 6%.

Of particular interest in the analysis of a plot such as that in Figure 8 are the relative frequencies of responses not classified alike by the latency and slope criteria. For the data presented in Figure 8, the incidence of responses categorized as conditioned by slope but voluntary by latency was originally low and not greatly affected by applying successively more strict criteria. On the other hand, the incidence of responses classified as conditioned by latency and voluntary by slope was initially about 27%, and it is of interest to determine what happens to this frequency as *S*s are discarded. Looking at the dashed lines connecting filled circles in Figure 8, it is clear that not until we apply very strict criteria indeed does any appreciable diminution in the relative frequency of such responses occur. To the extent that one wishes to rid the data of responses of this ambiguous type, the latency discard criterion leaves much to be desired.

In Figure 9 are presented comparable data for discard criteria based on the slopes of responses. As before, at the right end of the abscissa all *S*s would be discarded who had 0% or more responses with slopes steeper than 40%. At the left end of the abscissa only those would be discarded who had 100% responses with slopes greater than 40%. A gross comparison of the data in Figure 9 with the data in Figure 8 reveals at least two things. First, the cost in data thrown away of applying a given percentage discard criterion is greater with a slope criterion than with a latency criterion. For example, a 50% criterion value results in the discarding of some 40% of the *S*s with the slope criterion and only some 20% of the *S*s with a latency criterion. It also is apparent that with the greater cost comes a greater gain. Thus, with a 50% slope criterion the relative frequency of responses classified as voluntary by both slope and latency is reduced to below 3%. Even more important, the relative frequency of responses classified as conditioned by both of the criteria is increased to better than 80%. These values of 3% and 80% are to be compared with the values of 6% and 60% presented above in the case of the latency criterion. In addition, the ambiguous responses classified as voluntary by slope but conditioned by latency show a clear decline as the discard criterion becomes more strict.

It was shown by Hartman and Ross that the 300-msec. latency criterion and the 40% slope criterion were not equivalent bases for discarding *S*s from eyeblink conditioning experiments in which a ready signal was not employed. It is apparent from the data just discussed that the conventional latency and slope criteria are also not equivalent bases for discarding *S*s from at least some eyeblink conditioning experiments in which a ready signal is employed. As in the case of the Hartman and Ross data, the discrepancy between the two kinds of criteria arises because of the frequent occurrence of responses with steep slopes but with latencies considerably in excess of the original 300 msec. latency cutting point.

Before going on to analyze the implications of the discrepancies just noted, it may be of interest to look briefly at the learning curves generated by *S*s classified as voluntary or nonvoluntary by the conventional 50% latency and slope criteria. These data are presented in Figure 10. The 27 *S*s in Experiment 1 and 17 *S*s in Experiment 2 who were classified as nonvoluntary by both criteria show a gradually increasing learning curve beginning at approximately 12% conditioned responses and increasing to over 60% in Experiment 1 and about 55% in Experiment 2. Consistent with earlier reports (Goodrich, Markowitz, & Wall, 1963; Spence & Ross, 1959) that voluntary *S*s began at higher levels and conditioned to higher levels, the four *S*s in Experiment 1

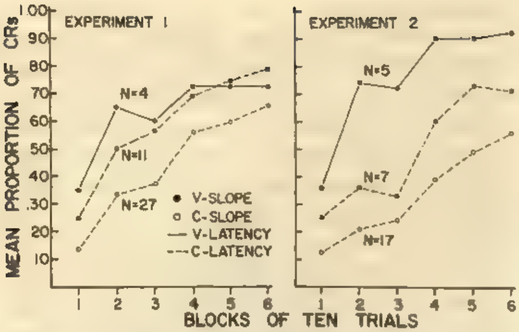

FIG. 10. Mean proportions of anticipatory responses in blocks of 10 trials for *S*s in both experiments falling in the categories formed by jointly classifying *S*s by the conventional slope and latency criteria.

and five in Experiment 2 who were classified as voluntary by both criteria demonstrated average learning curves consistently above the curves for the nonvoluntary *S*s. The remaining curve in each plot represents the 11 *S*s in Experiment 1 and the 7 in Experiment 2 who were classified as voluntary on the basis of their response slopes but as nonvoluntary on the basis of their latencies. It will be seen that although the relative position of the curves for these *S*s is not precisely the same in the two experiments, it falls between the other two functions. The intermediate position of these functions offers no clue as to the correct placement of these *S*s in the voluntary or nonvoluntary categories.

Instructions to Blink and Not Blink

The results of applying both latency and slope criteria to conditioning data were anything but completely satisfactory. Previous workers had evaluated the efficiency of the two criteria by determining to what extent they discriminated between responses *judged* to be either voluntary or conditioned. It was natural, then, that when confronted with the discrepancy between the two criteria discussed above, we should consider making judgments of the responses in our data. This required, of course, that we carefully study the judgmental criteria as listed by Spence and Ross. According to their report, in order to qualify as a V response an eyelid closure had to be *sharp, smooth, complete, and maintained* until after the air puff. The question arises as to the status of responses for which some of these four characteristics are true and others not. According to Spence and Ross, "Any response not meeting the voluntary form criteria was considered to be a CR [Spence & Ross, 1959, p. 378]." The two judges in the Spence and Ross experiment were able to agree fairly well on which responses were to be called voluntary, which conditioned, and which ambiguous, but each of the two judges placed approximately 20% of the responses in the questionable or ambiguous category. Apparently it was often difficult to decide whether a response was sharp, smooth, complete, and maintained. This difficulty became very clear when we attempted to apply the judgmental criteria to our data. A rather large number of responses appeared to be unclassifiable in terms of the judgmental criteria as we understood them. Moreover, it seemed clear that after applying the criteria we could not be sure that any differences between our results and those of Spence and Ross could be attributed to true differences in the data. They might just as well be attributable to differences in interpretation of the stated bases for making judgments. It was apparent that in a very real sense we did not know what we were doing. The results of applying the latency and slope criteria had led us to suspect that we were not sure these criteria were doing what they were supposed to do, and then we had discovered that we were not even sure what they were supposed to do.

To eliminate V responses and *S*s, we must be able to identify such responses and *S*s. The data discussed above were not consistent with the assumption that the conventional criteria were adequate. Thus it was decided to repeat and extend the kind of work which Spence and Taylor (1951) and Spence and Ross (1959) previously had reported rather casually. That is, we would seek examples of V and C responses and determine their differentiating descriptive characteristics, if any.

To obtain a pool of responses which could reasonably be regarded as voluntary, *S*s

were instructed to blink while the CS was being presented. For such Instructed-blink *Ss*, or at least for the subgroup of such *Ss* who received no air puff following the CS, it is reasonable to assume that no conditioning can occur.[7] To obtain a pool of responses which could reasonably be regarded as conditioned, other *Ss* were instructed not to blink while the CS was on. Although probably less reasonable than the analogous assumption with respect to the Instructed-blink *Ss*, it was assumed that the responses provided by the Instructed-inhibit *Ss* were nonvoluntary and representative of C responses. The plan, then, was to examine closely the slope and latency characteristics of the responses obtained under the two instruction conditions to determine descriptive characteristics distinguishing V from C responses.

Learning curves. Before turning to the latency and slope characteristics of the responses given by the two kinds of *Ss* just described, let us look briefly at the effect of the two kinds of instructions on the overall level of responding. Figure 11 contains the mean proportion of anticipatory responses in blocks of 10 trials for each of the six different conditions employed in the experiment. The right-hand panel shows conventional learning curves for the *Ss* run under standard conditioning instructions. At the top of the left-hand panel we see the corresponding data for *Ss* run under instructions to blink. Clearly, these *Ss* cooperated in complying with the instructions; during most of the session they responded on about 95% of the trials. In marked contrast to both these data and the data from the *Ss* run under standard conditioning instructions are the data in the lower left-hand portion of Figure 11 for the *Ss* instructed not to blink. Clearly, these *Ss* were able to

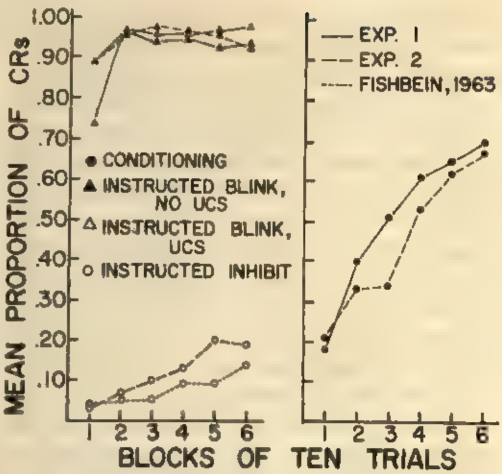

FIG. 11. Mean proportions of anticipatory responses in blocks of 10 trials for *Ss* in the several instruction conditions. At the right are the Conditioning *Ss*, 42 in Experiment 1 and 29 in Experiment 2. At the lower left are the Instructed-inhibit *Ss*, 30 in Fishbein's experiment and 52 in Experiment 2. At the upper left are the Instructed-blink *Ss*, 36 with UCS and 42 without the UCS in Experiment 1, and 18 without the UCS in Experiment 2.

inhibit responding to a considerable extent; even at the termination of 60 trials the group functions had not risen above 20% conditioned responses. The two curves, one from Fishbein's experiment and one from Experiment 2 are in close agreement. Their relative positions may be explained by the fact that a 2-psi air puff was employed in Fishbein's work, whereas a 5-psi puff was employed in Experiment 2. Presumably, responding based on a 5-psi puff would be more difficult to inhibit. The finding that instructions not to blink resulted in a response level markedly lower than that produced by standard conditioning instructions is consistent with earlier work on the effects of inhibitory instructions (e.g., Norris & Grant, 1948).

Latency and slope distributions. Figures 12 through 15 contain latency and slope distributions for the various instruction conditions. Figure 12 presents the frequency distributions of latency for the *Ss* who received instructions to blink while the CS was on. The functions for the three groups of *Ss* were very much alike, and the latency characteristics apparently changed very

[7] It has been brought to the writer's attention that Peak (1933) reported an unpublished study showing that for some *Ss* instructed to blink "...the secondary winking response became so automatic and habitual that even under conditions of...'no effort'...secondary reactions frequently occurred [p. 82]." Aside from our inability to evaluate the reliability of this finding, there is no information as to the nature of the obtained reactions. In any case, they did not result from CS-UCS pairings.

little between the first and second halves of the experiment. The mode of these distributions fell between 200 and 300 msec., consistent with other data obtained under similar instructions (Gormezano & Moore, 1962; Hartman, Grant, & Ross, 1960) and consistent with earlier work by Spence and Ross in which voluntary form responses obtained during conditioning occurred with a mode in this region. A comparison of the top two panels in Figure 12 shows that the presence or absence of the UCS apparently had little effect upon the latency distribution. The possible exception to this statement is the minor mode occurring in the vicinity of 125 msec. in the top panel under conditions in which the UCS was presented on each trial. This small distribution presumably represents unconditioned responses to the CS. The fact that such a distribution was more prominent when an air puff was presented may be explained by a sensitizing effect of the air puff.

Figure 13 presents the frequency distributions of response slopes for the same

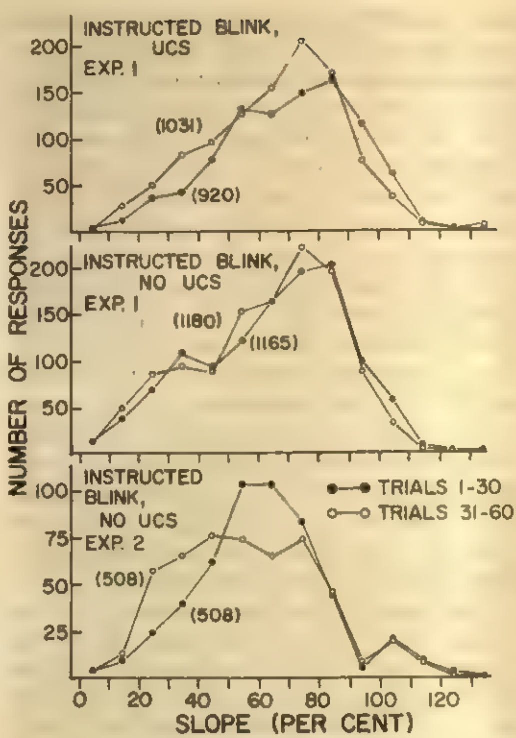

FIG. 13. Distributions of response slopes for the Instructed-blink *Ss* in the two experiments at two stages in training. Data include only responses with latencies greater than 150 msec. Numbers of *Ss* in the three panels are, from the top, 36, 42, and 18.

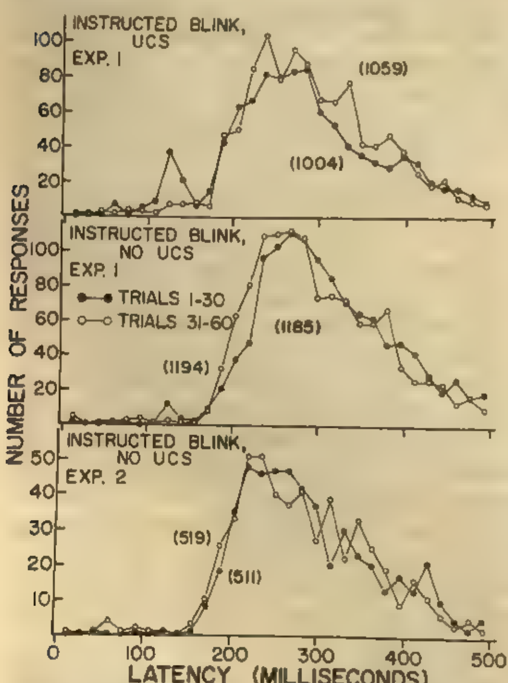

FIG. 12. Distributions of response latencies for the Instructed-blink *Ss* in the two experiments at two stages in training. The numbers of *Ss* in the three panels are, from the top, 36, 42, and 18.

three experimental conditions. The distributions in the top two panels, which represent the two Instructed-blink conditions from Experiment 1, are remarkably alike.[8] They

[8] It was noted that among the Instructed-blink *Ss*, those who received the UCS tended to have their eyes at least partially closed when the puff arrived. In contrast, those *Ss* who did not receive the UCS usually had their eyes open completely at the time a UCS would have arrived. Thus although the slope and latency distributions, as well as the overall frequency of responding, were very much alike for these subgroups, the "duration" of the response was different. This independence of duration from slope and latency raises the possibility that duration of V responses may not be a fruitful general basis for identifying such responses. As we have seen, latency distributions of presumed V responses have also proved labile under several variables (Goodrich, 1964a; Gormezano & Moore, 1962; Hartman, Grant, & Ross, 1960). If the *slope* of V responses should prove to be generally invariant, it would be the best candidate on these grounds for the role of identifying such responses.

both show a mode at approximately 75% slope, skewness to the left, and little systematic change between the first half and the second half of the experiment. In general, they appear to be unimodal. The lower panel contains the slope distributions of the *Ss* in Experiment 2 run under Instructed-blink conditions. Unaccountably, these distributions appear to differ from the corresponding distributions in the upper two panels. The mode occurs in the vicinity of 55% slope instead of 75% slope, and there appears to be a shift between the first half and the second half of the experiment toward more shallow responses.

Figures 14 and 15 present frequency distributions of latency and slope for the *Ss* in Fishbein's experiment and in Experiment 1 who were instructed *not* to blink while the CS was on. Figure 14 shows that latencies under these instruction conditions occurred predominantly late in the CS-US interval with a mode beyond 400 msec. There was a remarkable absence of responding prior to 300 msec., the point conventionally said to

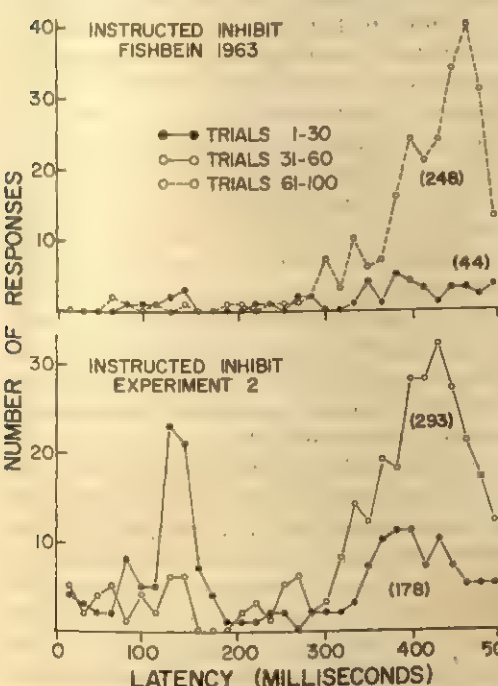

FIG. 14. Distributions of response latencies for the Instructed-inhibit *Ss* in Fishbein's experiment ($N = 30$) and Experiment 2 ($N = 52$) at two stages in training.

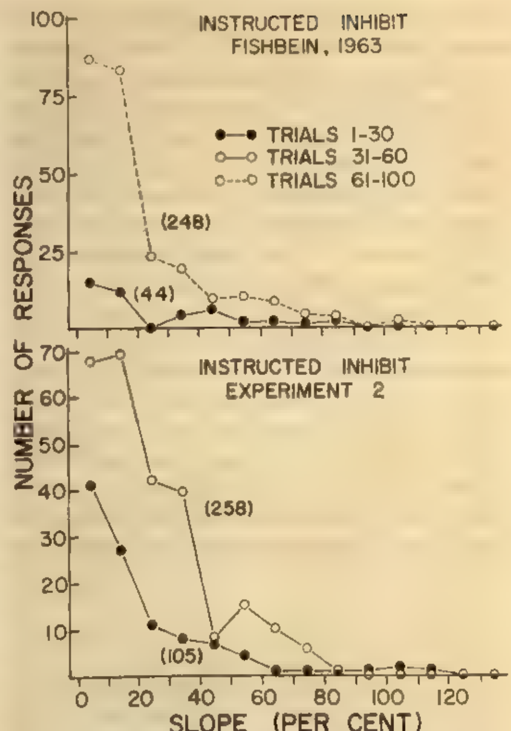

FIG. 15. Distributions of response slopes for the Instructed-inhibit *Ss* in Fishbein's experiment ($N = 30$) and Experiment 2 ($N = 52$) at two stages in training. Data in Experiment 2 include only responses with latencies greater than 150 msec. Fishbein's data include all latencies.

separate voluntary from conditioned responding. An exception to this generalization is contained in the results from Experiment 2 where a large secondary mode, presumably caused by unconditioned responses to the CS, occurred in the vicinity of 125 msec. The relative absence of this mode in Fishbein's experiment may be explained by his use of a relatively weaker air puff. As may be seen in Figure 15, the slopes of responses which occur under instructions not to blink are predominately shallow with a mode in the vicinity of 10%. Although responses with slopes less steep than 40% are in the majority, some responses occur with steeper slopes.

The distributions just presented in Figures 12–15 clearly do not suggest any unambiguous dichotomies for distinguishing V from C responses. The responses assumed to be voluntary, those obtained under in-

structions to blink, are distributed rather widely over the range of possible latencies and the range of possible slopes. Only by choosing very short latencies or very steep slopes could regions be defined in which the great preponderance of V responses fall. The situation is somewhat better in the case of responses assumed to be nonvoluntary, those obtained under instructions not to blink. In this case, the responses tend to be more concentrated along both the latency and the slope scales. The important consideration, however, is whether by jointly considering Instructed-blink and Instructed-inhibit conditions one can choose cutting points on the latency or slope scale which would permit unambiguous classification of responses as voluntary or conditioned. It is apparent from examining the data in Figures 12–15 that no such unambiguous cutting points exist.

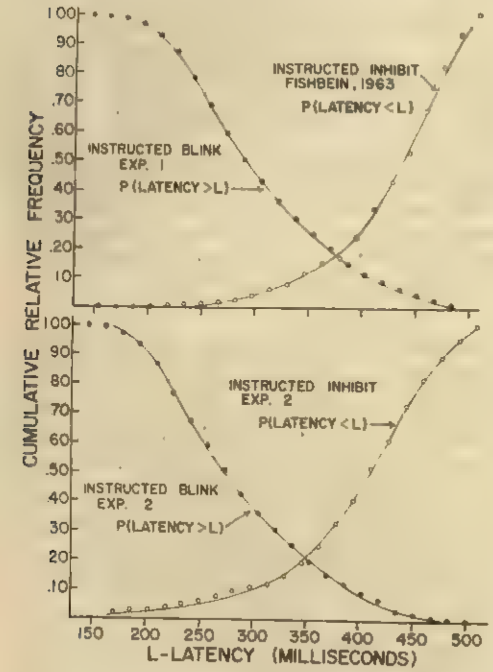

FIG. 16. Cumulative relative frequencies of response latencies for Instructed-blink (no UCS) and Instructed-inhibit *Ss*. Data were obtained from Trials 1–60 (Trials 1–100 for Fishbein's data) and include only responses with latencies greater than 150 msec. The total number of responses in the Instructed-blink and Instructed-inhibit distributions were 2345 and 375, respectively, in the top panel, and 1016 and 363, respectively, in the bottom panel.

Error probabilities in classifying responses. We may illustrate the problem of determining cutting points by plotting on the same graphs the data from Instructed-blink and Instructed-inhibit *Ss* in such a way as to represent the probabilities of erroneous classifications for each possible cutting point. Figure 16 is such a representation for the latency data. We have seen that responses obtained under Instructed-blink conditions typically have shorter latencies than those obtained under Instructed-inhibit conditions. Therefore, we seek a latency cutting point such that responses with shorter latency than that point will be called voluntary and those with longer latencies will be called conditioned.

Ideally, the probability α of erroneously labeling as conditioned a response which really is voluntary would be zero, as would the probability β of erroneously labeling as voluntary a response which really is conditioned. It is clear from the data presented in Figure 16 that this ideal cannot be met in our data: for no point along the latency scale are the ordinates of the decreasing function, representing the probability α, and the increasing function, representing the probability β, both zero.

Because α and β will both be nonzero, and because one probability tends to increase when the other decreases as we change the cutting point, it is clear that in choosing a cutting point we must consider the relative undesirability of the two kinds of error. Is it more undesirable to label as conditioned a true V response, or to label as voluntary a true C response? The first type of error means the inclusion of a response of a kind we do not want to study. The second type of error means the discarding of a datum which has cost both time and effort to obtain. Although it is clear that assigning weights to these two types of error is a difficult task and that the results may be debatable, it seems fairly clear to the writer that the first kind of error—including true V responses as C responses—must be regarded as the more serious kind. It follows

that the cutting point should be chosen in such a way that α is smaller than β.

An examination of Figure 16 will show that the conventional latency cutting point of 300 msec. does not meet the requirements we have just argued. Not only is α not smaller than β, but the magnitude of α is sizable. Only about 4% (Experiment 1) to 10% (Experiment 2) of the true C responses would erroneously be eliminated, whereas 38% (Experiment 2) to 46% (Experiment 1) of the true V responses would erroneously be retained.

How shall a cutting point be determined? The "decision rule" which is generally adopted in testing statistical hypotheses involves controlling the probability of one kind of error at an arbitrary small value and letting the other error probability take on whatever value is dictated by the methodology and true state of affairs. This is most clearly a satisfactory procedure when the error whose probability is controllable is markedly more serious than the other error.

If we regard the error of including a true V response as markedly more serious than the error of discarding a true C response, we may follow a procedure similar to that in hypothesis testing. Let us set α, the probability of including a V response, at some arbitrary but small value, say .05, and let β fall where it will. An analysis of the data in Figure 16, using both linear extrapolation and the curves fitted "by eye" to the data points in both experiments, reveals that a cutting point of about 435 msec. obtains for $\alpha = .05$. The corresponding value of β falls between .50 and .60, a value which is unfortunately high but which the logic of our decision rule dictates we must tolerate. In summary, then, if we assume that the instruction conditions have resulted in "true" C and "true" V responses, and that these are representative of such responses in conditioning situations, then the use of a latency cutting point of 435 msec. will result in including as C responses only about 5% of the true V responses, and the simultaneous elimination of about 55% of the true C responses.

The same logic as that just used to develop a cutting point for response *latency* also applies to developing a cutting point for response *slope*. In Figure 17 are presented the slope data in a form analogous to that of the latency data in Figure 16. In the case of slope, we have already seen that responses obtained under Instructed-blink conditions typically have greater slopes than those obtained under Instructed-inhibit conditions. Thus what is required is a cutting point on the slope dimension such that responses with slopes greater than that value will be called V responses and those with slopes less than that value will be called C responses. It is clear from Figure 17 that no cutting point exists such that both α, the probability of labeling as conditioned a true V response, and β, the probability of labeling as voluntary a true C response, will be zero. As in the case of latency, we must deal with the fact that both α and β will generally be nonzero and that one will increase as the other decreases.

Just as we examined the effect upon α and β of the conventional latency cutting point, we may also examine the corresponding effect of the conventional cutting point of 40% relative slope. Study of Figure 17 will show that use of a 40% cutting point would lead to approximately equal α and β probabilities. The probability of erroneously including a V response and the probability of erroneously discarding a C response are both about .18. Although these values are somewhat more reasonable than the corresponding values for the conventional latency cutting point, an α value as large as .18 is quite inconsistent with our argument that the cost of making this kind of error is markedly greater than the cost of making the other kind of error.

The data in Figure 17 were analyzed to determine the slope cutting point which corresponded to an α value of .05. Using both linear extrapolation and the smoothed curve for the data of both experiments, we arrived at a cutting point in the region of 20% slope. The corresponding value of β was about .32–.41. In summary, then, the present analysis suggests that the use of a slope cutting point of 20% will result in the inclusion in our data of only about 5%

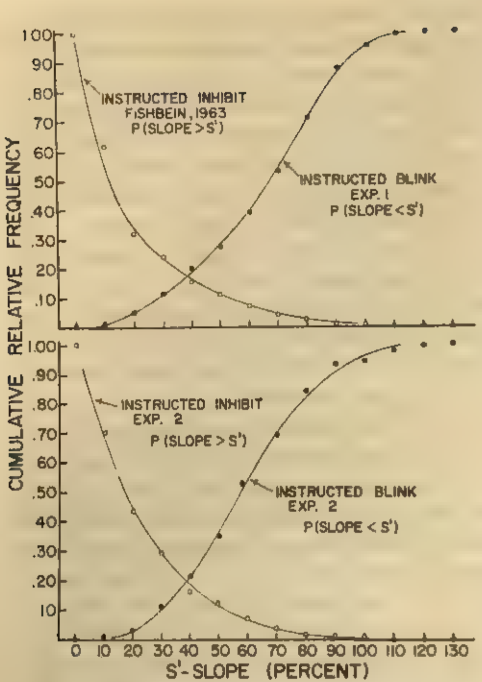

FIG. 17. Cumulative relative frequencies of response slopes for Instructed-blink (no UCS) and Instructed-inhibit *S*s. With the exception of Fishbein's experiment, data were obtained from Trials 1–60 and include only responses with latencies greater than 150 msec. Fishbein's data were obtained from Trials 1–100 and include all latencies. Numbers of responses are the same as those in Figure 16 except that Fishbein's data here consist of 392 responses.

of the true V responses, and the simultaneous elimination of about 36% of the true C responses.

Reanalysis of Conditioning Data with New Response Definitions

The analysis in the preceding section led to latency and slope cutting points which were quite different from the values which were recommended by previous research and which were employed in a previous section of this paper to analyze the relation between slope and latency in conditioning data. Figure 18 shows for each experiment the total response pool broken down into the four response categories formed by applying simultaneously the new slope and latency cutting points. The agreement between experiments is excellent. In both cases, more than half of all responses were classified as voluntary by both criteria, and fewer than 10% were classified as conditioned by both criteria. More responses were labeled voluntary than conditioned by both the slope and latency criteria, reversing the relation in the data presented earlier in Figure 7 for the conventional criteria. Whatever else the change in cutting points may have accomplished, it is clear that it did not eliminate the fact that slope and latency are not interchangeable ways of identifying V and C responses.

Distributions of V and C responses. Latency distributions for V and C responses defined by slope, and slope distributions for V and C responses defined by latency, were presented earlier in Figures 5 and 6 for the conventional cutting points. The corresponding distributions for the new cutting points will not be presented here because they do not take us much further than the summary data in Figure 18. As a result of basing the cutting points on considerations

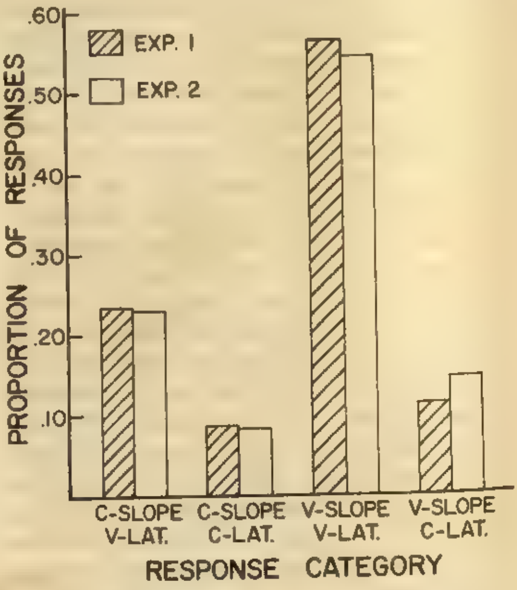

FIG. 18. Proportions of responses falling in the four categories formed by jointly classifying responses by the new slope and latency criteria. Data were obtained from the Conditioning *S*s in Experiment 1 ($N = 42$) and Experiment 2 ($N = 29$) over Trials 1–60. Only responses with latencies greater than 150 msec. were included, 1235 in Experiment 1 and 745 in Experiment 2.

of errors in classification, we know that such distributions contain certain proportions of erroneously classified responses. However, because these proportions refer to the true, rather than the obtained, numbers of V and C responses, the obtained distributions are distorted in a fashion not easily corrected.

Estimating the number of true V and C responses. Given our procedure for determining the latency and slope cutting points and the assumption of stability in the values of α and β for different sets of data, it becomes possible to estimate for any collection of conditioning data the true numbers of C and V responses.

Recall that α is the probability of including a true V response and β is the probability of discarding a true C response. Now let C_0 and V_0 represent the *obtained* numbers of included (C) responses and discarded (V) responses, respectively, and C_t and V_t represent the *true* numbers of C and V responses, respectively. Then $C_0 = \alpha V_t + (1 - \beta)C_t$ and $V_0 = (1 - \alpha)V_t + \beta C_t$.

Because numerical values of α and β arise from the procedures used in differentiating the responses into a retained set consisting of exactly C_0 responses and a discarded set consisting of exactly V_0 responses, there are only two unknowns in the two equations above, C_t and V_t. Thus solutions may be found for C_t and V_t. For any set of conditioning data, then, we not only can partition the total pool of responses into two sets, one of which theoretically contains all C responses except for $100\alpha\%$ of the true number of V responses, but we may also obtain an estimate of the true number of V (and C) responses. In Experiment 1, the estimated proportion of all responses which were really V responses was .63 using the latency cutting point and .32 using the slope cutting point. In Experiment 2 the corresponding values were .56 and .35. It is apparent that the results of the two experiments are in good agreement, whereas the slope and latency procedures lead to rather disparate outcomes.

Before gaining too much satisfaction from the procedures just described, we must recall from our earlier discussion that the discarding of responses per se is not generally a meaningful tool of analysis in studies of eyeblink conditioning. For similar reasons, the possibility of estimating how many of the responses obtained in an experiment are actually V responses will not be regarded as particularly useful, since it permits no way of obtaining an estimate of the probability of the occurrence of a C response unbiased by the occurrence of a V response. Nonetheless, it may be of some interest to use estimates of V_t and C_t as dependent variables in experiments which involve the manipulation of variables, such as UCS intensity, which presumably influence the adoption by *S* of a voluntary mode of responding.

In any case, it is possible to show that the equations above impose theoretical constraints on any set of obtained data. These restraints provide a way of testing the procedures and assumptions employed here. We note that the maximum number of responses will be discarded when every response is really voluntary, and the minimum number of responses will be discarded when every response is really conditioned. If all responses were voluntary, then $C_t = 0$ so that $C_0 = \alpha V_t$ and $V_0 = (1 - \alpha)V_t$. Thus when all responses are voluntary, the relative frequency of discarded responses would be $(1 - \alpha)$. When some of the responses are actually conditioned, the relative frequency of discarded responses would be less than $(1 - \alpha)$. Similarly, if all of the responses were actually conditioned, then $V_t = 0$ so that $C_0 = (1 - \beta)C_t$ and $V_0 = \beta C_t$. The relative frequency of discarded responses under these conditions would be β. To summarize these implications of the present model, the smallest possible relative frequency of discarded responses will be β, obtained when all of the responses are actually conditioned, and the largest possible relative frequency of discarded responses will be $(1 - \alpha)$, obtained when all of the responses are actually voluntary.

To illustrate a possible application of these deductions, let us presume that airpuff intensity is directly related to the adopting of the voluntary mode of responding. We would predict from the present ar-

guments that no matter how weak the puff, no fewer than 100β% of the responses would be discarded as voluntary, and that no matter how intense the air puff no more than 100 $(1 - \alpha)$ percent would be so discarded. On the basis of our earlier findings, let us take α as .05 and β as .55 for latency and .36 for slope, so that the limiting percentages are 55% and 95% for latency and 36% and 95% for slope. The two sets of conditioning data reported here provide results which are consistent with these limits: 80% of the responses in Experiment 1 and 78% in Experiment 2 were discarded with the latency cutting point, and 68% and 69% with the slope cutting point.

Predicting the obtained frequency distributions. The assumptions made in developing the procedures described above suggest another possibility which can be used to test the adequacy of these assumptions. We have assumed that the Instructed-blink condition provides a picture of the distributions of true V responses, and the Instructed-inhibit condition provides a picture of the distributions of true C responses. In the preceding section we derived estimates of the relative frequency of true V responses in the conditioning data. Putting these relations together, it should be possible to predict the obtained slope and latency distributions for the Conditioning groups by combining the appropriate distributions from Instructed-blink and Instructed-inhibit conditions in the proportions dictated by the estimated relative frequency of true V and C responses in the data.

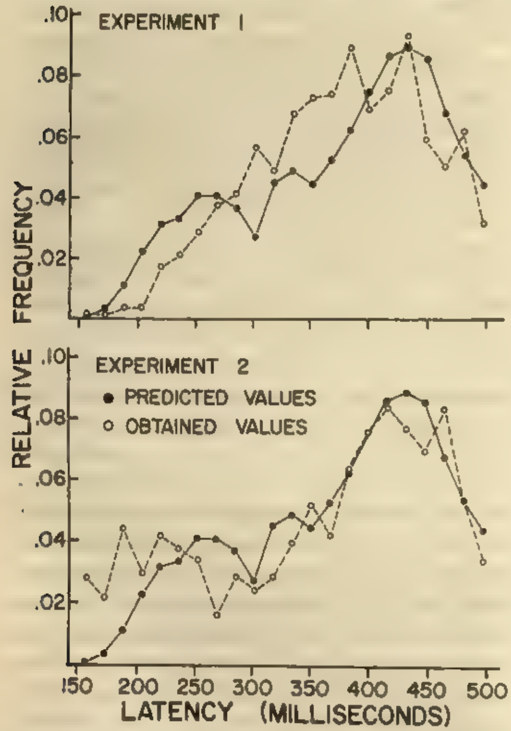

FIG. 19. Predicted and obtained latency distributions of conditioning data. Predicted values were obtained from a slope-derived estimate of the relative frequency of V responses (see text). Obtained data are based on Trials 31–60 for Conditioning *Ss* and contain only responses with latencies greater than 150 msec., 801 in Experiment 1 and 546 in Experiment 2.

Such derivations of the Conditioning data were carried out by first pooling across the two experiments the corresponding distributions for Instructed-blink *Ss* and for Instructed-inhibit *Ss*. Then the Instructed-blink and Instructed-inhibit distributions were combined in "correct" proportions to predict the Conditioning data. The "correct" proportions for one set of results were obtained by using the estimate of the proportion of true V responses which resulted from dichotomizing responses with a latency cutting point. We saw in the previous section that this proportion was .63 for Experiment 1 and .56 for Experiment 2. For the present purposes, a value of .60 was employed. Thus the resulting synthesized distributions of slope and latency contained 60% Instructed-blink responses and 40% Instructed-inhibit responses. Similarly, a single estimated proportion of true V responses, .33, was derived from the values in Experiment 1 (.32) and Experiment 2 (.35) which resulted from dichotomizing responses with a slope cutting point. In this case the synthesized distributions of slope and latency contained 33% Instructed-blink responses and 67% Instructed-inhibit responses.

The results of these analyses are shown in Figures 19–22. Figure 19 shows the predicted latency distributions in comparison with the actual obtained distributions for the Conditoning groups in the two experi-

ments. Both distributions have unit area. The predicted values contain 33% Instructed-blink responses. Analogous results are shown in Figure 20 for predicted values containing 60% Instructed-blink responses. It is apparent that the 33% value, which was determined from a slope cutting point, led to a more satisfactory fit to the obtained data in both experiments. The 60% value, determined from a latency cutting point, resulted in a poor fit caused by relatively too many predicted responses with short latency.

Figures 21 and 22 present the same kind of data for distributions of response slope. The predicted distributions in Figure 21 contain 33% Instructed-blink responses, whereas those in Figure 22 contain 60% such responses. For Experiment 1 the 33% value, derived from using the slope cutting point, again produces the better fit of predicted to obtained distributions. For Experiment 2 neither the 33% nor 60% values clearly leads to better prediction.

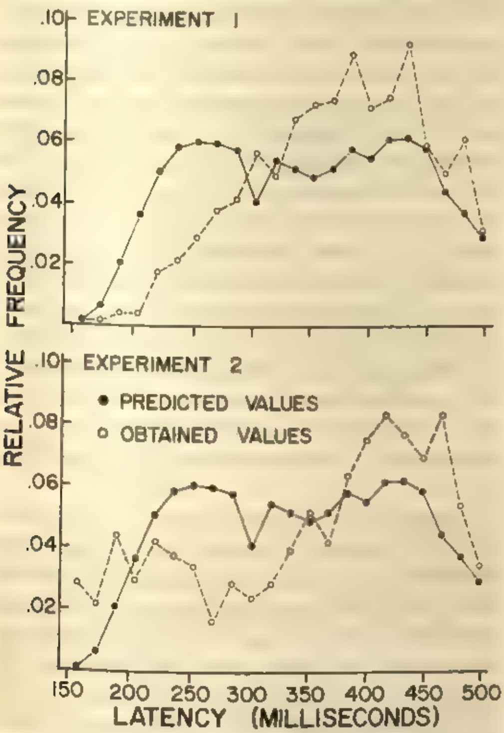

FIG. 20. Predicted and obtained latency distributions of conditioning data. Predicted values were obtained from a latency-derived estimate of the relative frequency of V responses (see text). Obtained data are identical with those in Figure 19.

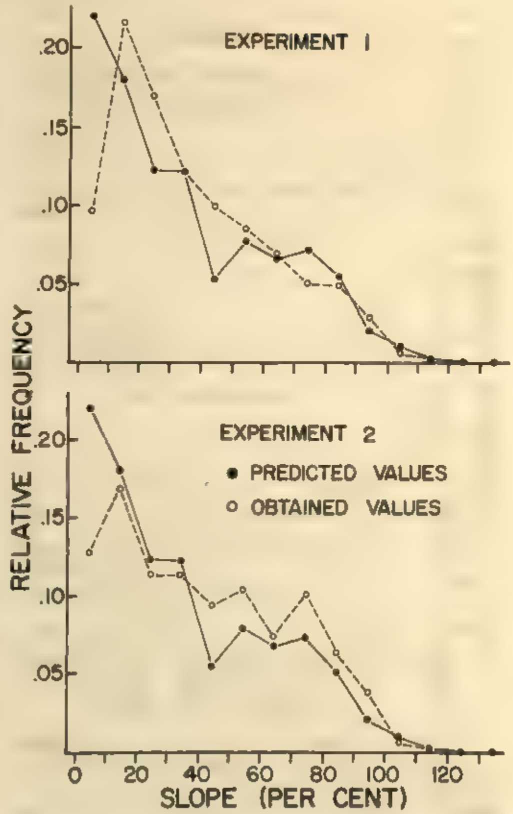

FIG. 21. Predicted and obtained slope distributions of conditioning data. Predicted values were obtained from a slope-derived estimate of the relative frequency of V responses (see text). Obtained data are based on Trials 31–60 for Conditioning *Ss* and contain only responses with latencies greater than 150 msec., 801 in Experiment 1 and 546 in Experiment 2.

The differences above between the 33% results and the 60% results are not the first instances in the present report of disparate results of applying methods based on slope and latency. All of the cases we have encountered have in common that many responses are labeled as voluntary by one criterion and conditioned by the other. In the previous cases, there was no basis within the experiment for preferring one cutting point over the other. In the present instance, differences are again apparent between outcomes based on slope and latency cutting points. This time, however, the dif-

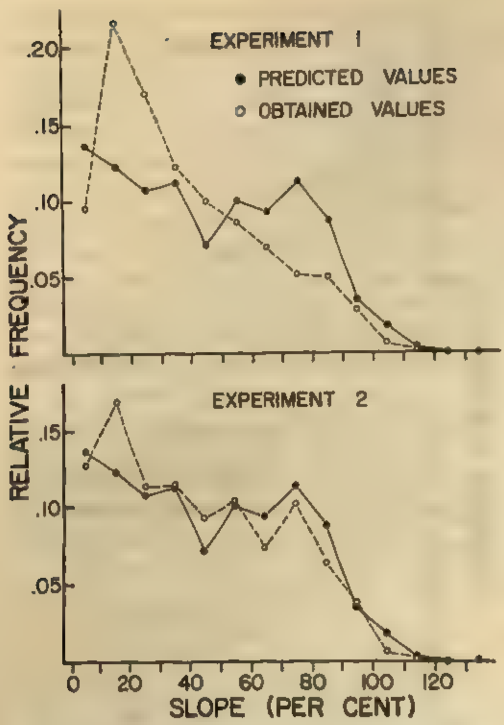

FIG. 22. Predicted and obtained slope distributions of conditioning data. Predicted values were obtained from a latency-derived estimate of the relative frequency of V responses (see text). Obtained data are identical with those in Figure 21.

ferences pertain to predictive accuracy and thus provide some grounds for choosing one kind of cutting point over the other. Overall, the data in Figures 19–22 suggest that the more accurate synthesis of the conditioning data results from combining Instructed-blink and Instructed-inhibit distributions in the proportions dictated by a *slope*, rather than latency, cutting point.

No attempt is made here to assess quantitatively the fit between predicted and obtained distributions. The usual methods are inappropriate because the distributions are pooled over both *Ss* and responses within *Ss*, and the intent of presenting these data is to reveal possibilities and problems, not to provide precise tests.

Effects of discarding V Ss. Entering into the preceding analyses have been the data from all *Ss* in the Conditioning groups. As we have already indicated, the practical value of identifying V responses is small unless this identification is used as a basis of discarding all the data from selected *Ss* in the experiment. The data in Figures 8 and 9, discussed earlier, showed what happened to the size and composition of the set of data remaining after discarding *Ss* as "voluntary responders." The V and C responses were defined by what we have called the conventional cutting points. We turn now to similar analyses carried out with the new cutting points determined by a consideration of errors in classification.

Figure 23 shows the effects upon the data which remain after *Ss* are discarded for meeting successively more stringent latency discard criteria. Towards the left end of each abscissa are criteria which eliminate only those *Ss* with high relative frequencies of V-latency responses. Toward the right end large numbers of *Ss* are eliminated as more and more of them meet the requirement of small proportions of V-latency responses. Study of these data shows that no criterion seems to even moderately well approximate the ideal state of affairs in which all but the responses labeled as conditioned by both slope and latency cutting points would be eliminated and a reasonable proportion of the data would be retained. As an illustration, we note that roughly half of the data has already been discarded with an 80% criterion, and that at that point the responses consistently labeled conditioned are still the smallest category and those consistently labeled voluntary are still present in large numbers. Even by tolerating unreasonably large losses of data, the remaining sample is only poorly rid of V responses by a latency criterion for discarding *Ss*.

Figure 24 shows the effects of discarding *Ss* on the basis of slope criteria. Comparison of the top portions of Figures 23 and 24 will show that for the same percentage discard criterion fewer *Ss* are eliminated with the slope criterion than with the latency criterion. This finding reflects the fact, already pointed out, that the slope cutting point identifies fewer responses as voluntary than does the latency cutting point. Study of the data in the lower portion of Figure 24 shows that discarding *Ss* by slope criteria is

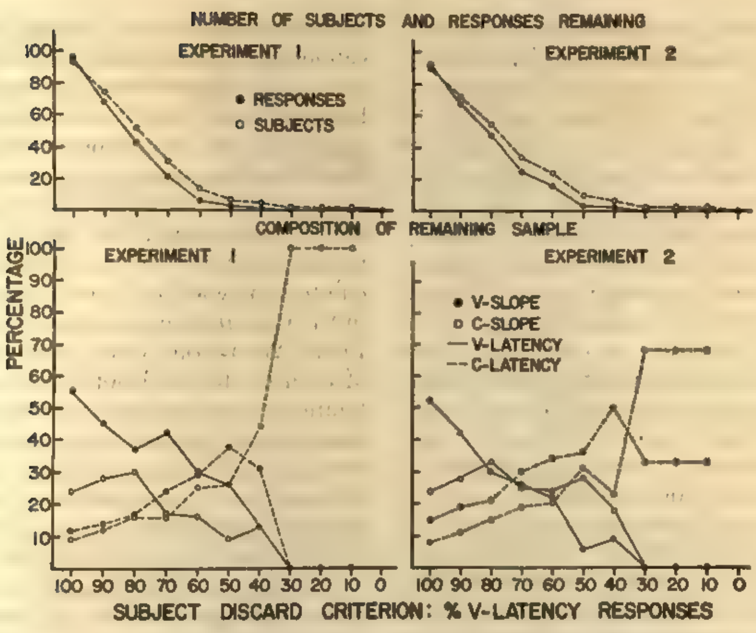

FIG. 23. Effects of discarding Conditioning *S*s with the modified latency procedure upon the amount and composition of data remaining. Data were obtained over Trials 1–60 and include only responses with latencies greater than 150 msec., 1235 in Experiment 1 and 745 in Experiment 2.

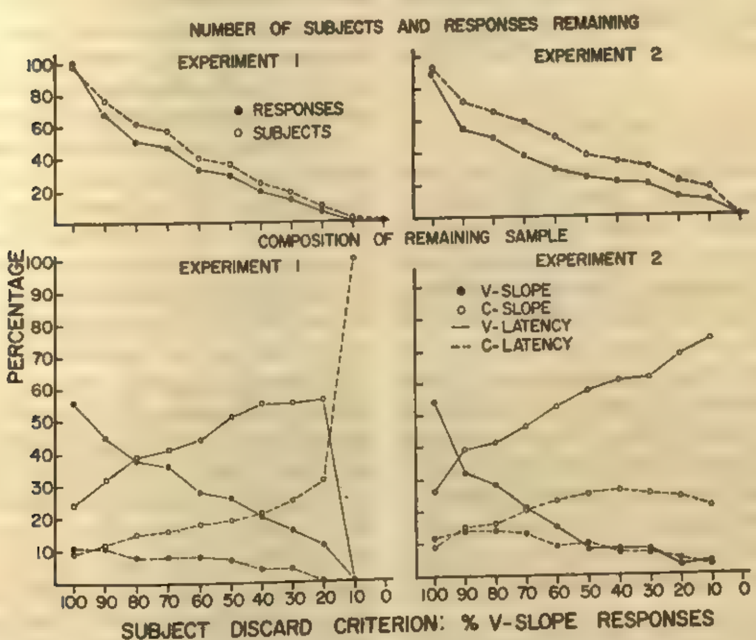

FIG. 24. Effects of discarding Conditioning *S*s with the modified slope procedure upon the amount and composition of data remaining. Data were obtained over Trials 1–60 and include only responses with latencies greater than 150 msec., 1235 in Experiment 1 and 745 in Experiment 2.

somewhat more satisfactory than discarding them by latency, as judged by the composition of the remaining data. Nonetheless, the slope criterion still leaves much to be desired. For example, half of the data are again lost when the criterion is no more severe than about 80%. At this point V-latency responses predominate over C-latency responses, and responses labeled consistently as voluntary greatly exceed in number those labeled consistently as conditioned. The situation is only moderately improved by tolerating larger losses in data.

Overall, neither the slope nor the latency criteria for discarding *S*s is particularly successful at achieving what it was designed to do: to discard from conditioning experiments a reasonable number of *S*s in order to rid the remaining data of V responses, defined by the new cutting points developed in this paper. The reader will recall that similar procedures were more satisfactory for the task of ridding the data of responses defined by the conventional cutting points. An improved method of classifying responses seems to have led to a markedly lowered efficiency of the procedures on which many workers in the area of eyeblink conditioning have depended for eliminating voluntary influences from their data.

A Direct Method of Identifying V Ss.

The methodology discussed so far has followed the earlier work (Hartman & Ross, 1961; Spence & Ross, 1959) by dealing with response distributions pooled over all *S*s and responses. Although responses are the units within these distributions, the individual *S*, rather than the individual response, is the actual *sampling* unit. If the latencies or slopes of responses of individual *S*s tend to cluster closely about average values, and the average values for different *S*s are spread over the range of possible values, the adding or subtracting of a few *S*s at random may have a marked effect on the shapes of the pooled response distributions.

The empirical question here is whether the shapes of obtained response distributions arise in large measure from the distribution of *S*s per se. Clearly, the question cannot be approached by asking about the shapes of distributions for individual *S*s, because an *S* could make at most 60 responses in the experiments reported here. The remaining approach involves looking at the way in which *S*s themselves are distributed. To this end, the median slope and median latency over all 60 trials were calculated for each *S*. These data for the *S*s in the Conditioning groups are presented in Figure 25. In the lower right-hand corners of both the top and bottom sections of this figure are scatter plots showing the relation between median slope and median latency, each point locating one *S*. In the upper portion of each section is the marginal latency distribution; in the left portion is the marginal slope distribution.

An examination of Figure 25 in conjunction with Figures 3 and 4, discussed previously, will show that the distributions of *S*s' medians have roughly the same shapes as the pooled response distributions. Consider, for example, the latency distributions for Experiment 2. Both the response and *S* distributions are skewed to the left. Moreover, the presence of a mode around 200 msec. in the response distribution, which did not appear in Experiment 1, is paralleled (and probably explained) by the left-hand mode in the distribution of *S*s' medians.

Figure 26 presents distributions of individual *S*s along the latency and slope dimensions for *S*s run under the Instructed-inhibit and Instructed-blink conditions. It is apparent from both the scatter plots and the marginal distributions that these distributions of *S*s' medians also correspond rather well to the distributions in Figures 12–15 based on pooled responses. Along the latency dimension, *S*s instructed to blink tended to fall in the region between 200 and 400 msec. with a mode near the middle of the range, whereas *S*s run under instructions not to blink fell mostly in the region between 350 and 500 msec. Along the slope dimension, *S*s run under instructions to blink were distributed in the region of steep slopes whereas *S*s run under instructions not to blink were located generally in the region of rather shallow slopes.

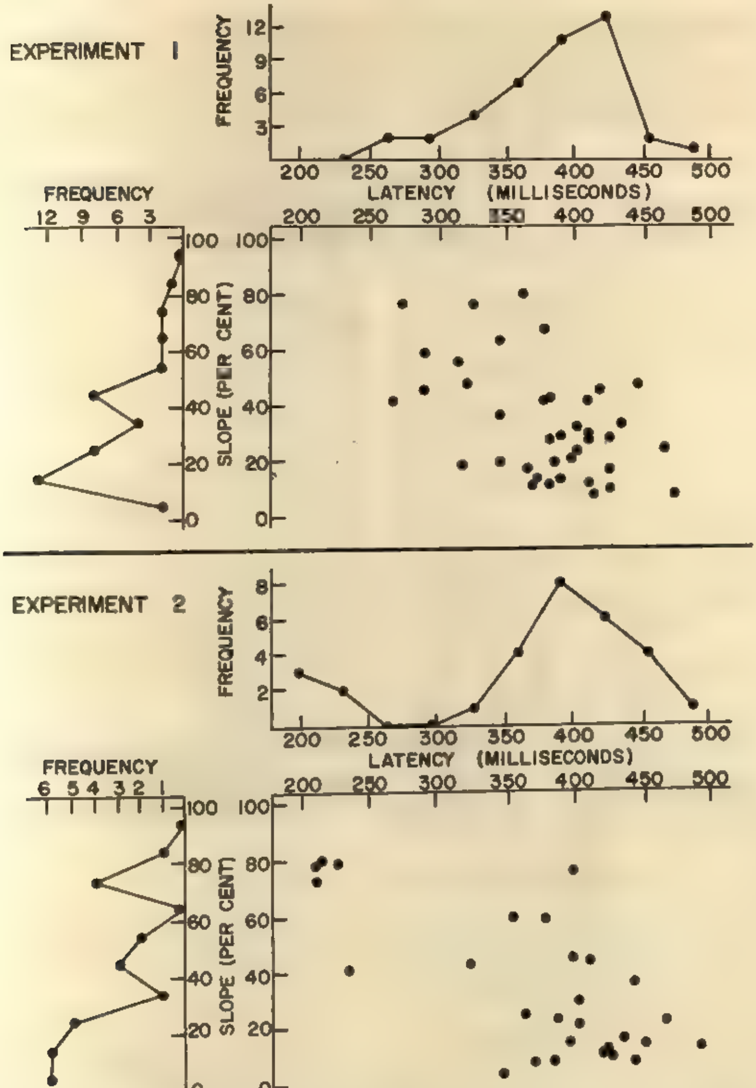

FIG. 25. Scatter plots and marginal distributions of median slopes and median latencies for individual Conditioning *S*s. Data were obtained over Trials 1–60 and include only responses with latencies greater than 150 msec.

In summary, it seems clear that in large measure the shapes of the pooled response distributions discussed in previous work and in earlier sections of this report reflect the way in which the typical responses of individual *S*s are distributed, not simply the processes occuring in each *S*.

Error probabilities in classifying Ss. Attention to the data of individual *S*s suggests a redefinition of the task of identifying V and C *S*s. Thus far the procedure for discarding *S*s has started with first classifying responses, then discarding *S*s. The validation of the discarding of *S*s consisted in evaluating how successfully it eliminated responses of certain descriptions from the data. Apart from possible complications arising from the fact that the classification of responses rests on pooled response distributions, we have seen above that the conventional procedure simply does not work well.

Perhaps a more direct approach is possible, one in which *S*s would directly be clas-

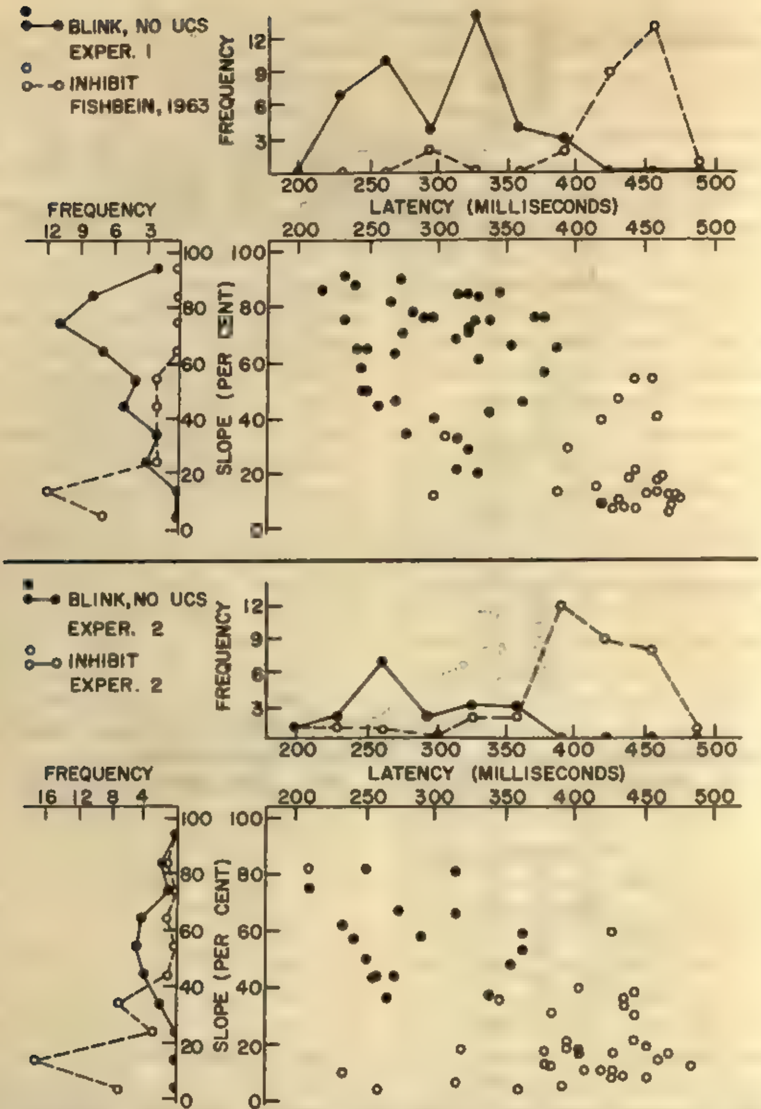

FIG. 26. Scatter plots and marginal distributions of median slopes and median latencies for individual *S*s in the Instructed-inhibit and Instructed-blink (no UCS) conditions. Data were obtained over Trials 1–60 and include only responses with latencies greater than 150 msec.

sified as voluntary or not. With such a procedure, the experimental conditions involving blink and inhibit instructions in this paper would be regarded as sources of *subject* types rather than *response* types. The logic formerly applied to error probabilities in classifying responses would here be applied to the analogous error probabilities in classifying *S*s.

The α and β error probabilities in attempting to classify *S*s directly are shown in Figures 27 and 28. As before, it is assumed that instructions to blink provide V *S*s, and instructions not to blink provide C *S*s. The ordinates of the descending functions in Figure 27 show values of α, the probability of including a V *S*. The ordinates of the ascending functions are values of β, the probability of discarding a C *S*. We must keep in mind in the following discussion that these distributions of *S*s' medians are less stable than the analogous

plots, discussed earlier, of distributions based on pooled responses. In spite of the probable limitations imposed by the small samples involved here, it is possible to show how one might proceed to develop cutting points for classifying *S*s. The logic is exactly like that previously employed. Because the cost of erroneously including a V *S* is much greater, it is supposed, than the cost of discarding a C *S*, we set α at the small arbitrary value of .05. The latency cutting point which corresponds to $\alpha = .05$ is roughly 380 msec. Thus if we discard an *S* whenever his median latency is less than 380 msec., we will in the long run include only about 5% of the true V *S*s in our sample. Depending on how we use Figure 27 to estimate β, we should expect to discard between 10% and 30% of the true C *S*s.

Figure 28 represents the data for determining a slope cutting point. Here α is represented by the ordinates of the ascending functions, and β by the ordinates of the descending functions. The slope value which results from setting α at .05 is roughly 30%. The corresponding value of β is somewhere between .15 and .24. To state again the implications of this analysis: using a slope cutting point of 30% should result in including about 5% of the true V *S*s and discarding 15% to 24% of the true C *S*s.

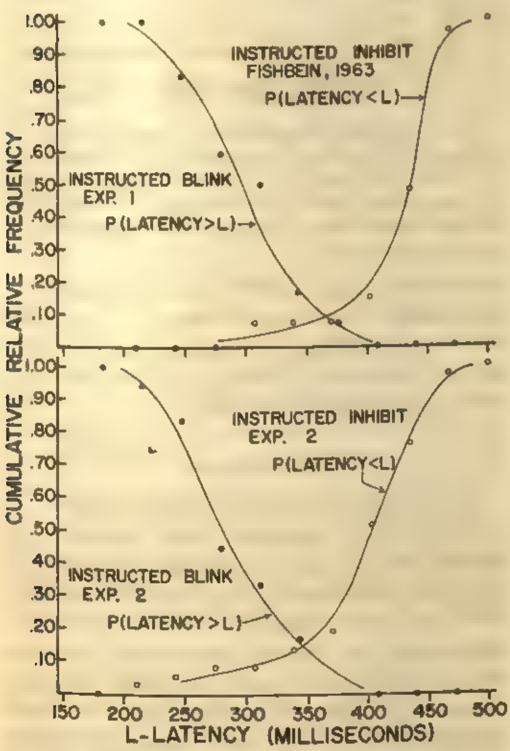

FIG. 27. Cumulative relative frequencies of median response latencies for individual *S*s in the Instructed-inhibit and Instructed-blink conditions.

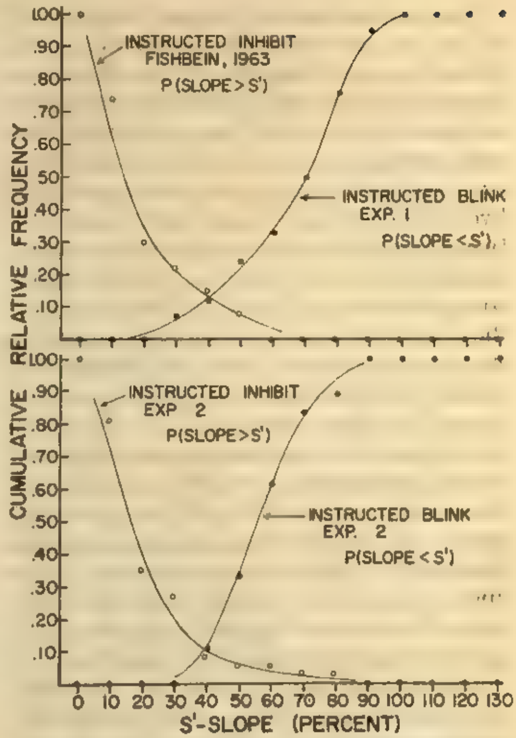

FIG. 28. Cumulative relative frequencies of median response slopes for individual *S*s in the Instructed-inhibit and Instructed-blink conditions.

It is interesting to note that when *both* the latency and slope cutting points are applied simultaneously in a disjunctive fashion, *none* of the Instructed-blink *S*s are classified as C *S*s and 46% of the Instructed-inhibit *S*s are classified as V *S*s. That is, if an *S* is discarded whenever *either* his median latency is less than 380 msec. *or* his median slope is greater than 30%, very nearly 100% of the true V *S*s will be discarded, whereas 54% of the true C *S*s are retained. (Some V *S*s undoubtedly would be retained, in spite of the fact that all of the present Instructed-blink *S*s would be discarded.)

The breakdown of the total number of *S*s in the two Conditioning groups in the present study into the four categories formed by slope and latency cutting points is shown in Figure 29. The results of the two experiments are in good agreement. However, we again encounter a discrepancy between latency and slope procedures. The two categories containing *S*s which were not classified alike by the two cutting points contain about 25% of all *S*s. In the light of the results discussed in the last paragraph showing the efficacy of a disjunctive criterion, the discrepancy between slope and latency results illustrated in Figure 29 would probably best be circumvented with the disjunctive criterion. With this criterion about 46% of the *S*s would be discarded, which admittedly is not a happy outcome. Nonetheless, if the assumptions are valid under which the cutting points were developed, the remaining 54% of the *S*s should consist almost exclusively of true C *S*s.

Learning curves for V and C Ss. As a final order of business, Figure 30 presents the learning curves of the *S*s who did and did not meet the disjunctive criterion above. The discarded *S*s responded more frequently throughout training than the retained *S*s. These data are similar to those cited by

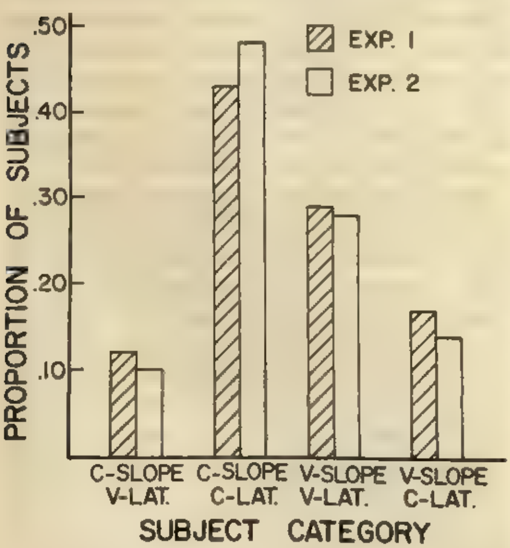

FIG. 29. Proportions of Conditioning *S*s falling in the four categories formed by jointly classifying *S*s with the direct latency and slope criteria.

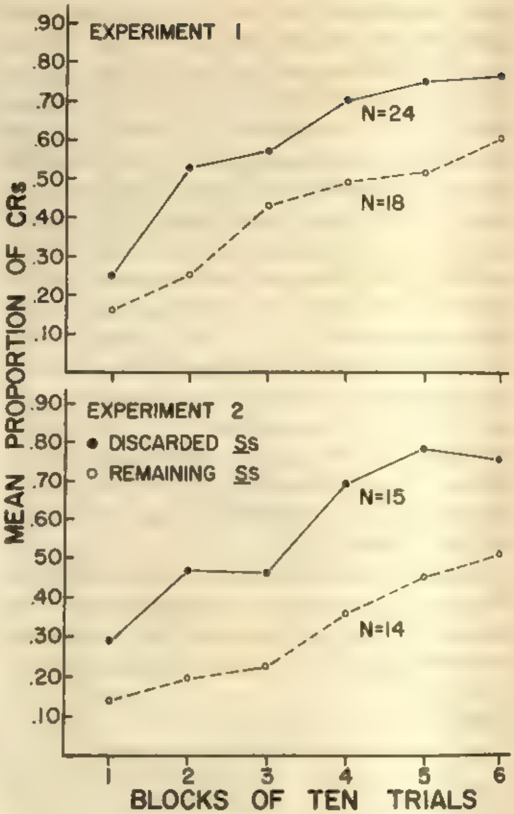

FIG. 30. Mean proportions of anticipatory responses in blocks of 10 trials for Conditioning *S*s who were retained as C *S*s by both slope and latency direct criteria (remaining *S*s) and who failed to meet both of these criteria (discarded *S*s).

Spence and Ross (1959) and by Goodrich, Markowitz, and Wall (1963). It is clear that the estimated probability of an anticipatory response is smaller after the discarding of suspected V *S*s.

DISCUSSION

The major functions of the present report are two in number: first, to display in greater detail than heretofore the data of eyeblink conditioning experiments as these data have been analyzed for V responses, and second, to repeat and extend the approach originated by Spence and his associates to attempt an evaluation of the current status of this methodology. The initial examination of the conditioning data suggested that even samples as large as those employed here are insufficient to insure

that pooled distributions will have the same shapes upon replication. This is but another instance of the extremely large variability often characteristic of eyeblink conditioning data (cf. Spence, 1964, pp. 131, 133). In addition, the present results show clearly that the discrepancy between latency and slope analyses of V responses is not found only when a ready signal is omitted; the two modes of analysis were not equivalent methods of discarding V *S*s when a ready signal *was* employed.

One outcome of the present extension of the approach of Spence and Taylor (1951) and Spence and Ross (1959) was the development of new cutting points for identifying V responses, cutting points which were explicitly rationalized in terms of the probabilities of errors of classification. The deductive consequences which arose from the assumptions used in developing these cutting points cannot be regarded as either supporting or casting doubt on the assumptions. It remains to be seen whether future research will be able to make use of the suggested methods. Unfortunately, use of the cutting points for identifying V responses to eliminate V *S*s in the manner of Spence and Ross (1959) was quite unsuccessful and casts doubt on this particular approach to discarding *S*s.

The most interesting result of extending the basic approach of earlier investigations was the direct identification of V *S*s by their resemblance to *S*s instructed to blink. The principal recommendation arising from this procedure and the results presented here is that an *S* be eliminated from an eyeblink conditioning experiment if his median response latency is less than 380 msec. or his median relative response slope exceeds 30%. The main rationale for this discard rule is the implication of the present experiments that nearly all of the V *S*s would be eliminated by application of this disjunctive rule.

Unfortunately this rationale is not fully adequate. First, the distributions of *S*s' medians upon which the cutting points were based contained too few cases to permit uncritical application of the exact values reported here. Clearly, validation research is required. Second, it must be remembered that a particular set of experimental parameters was employed in the present experiments. Certain of the numerical results may not be invariant with changes in these parameters. One may entertain some cautious optimism, however, in view of the finding that latency and slope distributions for *S*s instructed to blink were invariant with presence-absence of the UCS (Experiment 1, Figures 12 and 13). Third, we cannot be altogether happy with the assumptions concerning the relation between instructions to blink or not blink and the occurrence of V and C responses. Few would argue that the description of *S*s instructed to blink as "voluntary" is inappropriate. But we cannot yet be sure that "voluntary" means the same thing here as in the conditioning situation. Does, for example, a V *S* whose responses have as their controlling outcome the successful compliance with instructions to blink have the same median slope and latency as an *S* whose responses have as their controlling outcome the mitigation of the unpleasantness of a puff in the eye? We have assumed the answer is "yes," but the question has not been put to an adequate test.

More open to question is the assumption that responses which occur under instructions not to blink are conditioned or nonvoluntary. It does appear reasonable that such responses are not voluntary *blinks*. But are they pure conditioned responses? The possibility exists that such responses are a complicated resultant of a true conditioned closure with a *voluntary opening*. It is interesting to note that the typical response under inhibit conditions, like the typical conditioned response described by Spence and Ross (1959, p. 378), was a gradual and irregular movement resembling a tense, slow, and controlled limb movement. The existence of an eye-opening response cannot be doubted. Record C in Figure 1 illustrates a trial on which an opening occurred without a closure. Such responses were not uncommon under the inhibit condition and are occasionally seen in *S*s run under ordinary conditioning instructions. If the form of the typical con-

ditioned response is thought to be the resultant of opposing blink and opening responses, it is apparent that we may not have isolated the conditioned *blink* at all.

Fortunately, the rationale for the procedures recommended in this report depends far less critically on interpreting the results of instructions not to blink than it does on interpreting the results of instructions to blink. This difference in importance arises because of the presumption that an error of falsely including a V *S* is far more serious than the error of falsely discarding a C *S*. Thus the only matter which actually depends upon interpreting the distributions from *S*s instructed not to blink is the estimation of the proportion of true C *S*s which are discarded.

One additional problem may be raised. An *S* under conditioning instructions may adopt, or even abandon, the voluntary mode of responding any time during a series of trials. The available procedures for discarding V *S*s, including those developed in the present paper, make use of the median of *S*'s latencies or slopes over the entire training session. It is clear that if *S* became a voluntary responder during the second half of the session, he probably would not be discarded and his data for the latter part of acquisition would be combined with data from C *S*s. A check was made for the two sets of conditioning data in the present study of whether different *S*s were identified as V *S*s when all 60 trials were included as compared with when only the last 30 trials were included. A few *S*s changed classification, but because these generally were *S*s who made few responses and because the number of cases on which the medians were based was not generally the same for Trials 1–30 as for Trials 31–60, a conclusion could not be reached as to whether *S*s had actually adopted the voluntary mode of responding late in training. The problem may be more serious with weaker UCS intensities than with the strong 5-psi puff employed here because a weak UCS may become increasingly aversive with successive trials. A very strong UCS, in contrast, may induce avoidance from the outset.

Clearly the present work has not established that rules for identifying voluntary processes in eyeblink conditioning are necessary. Gormezano (e.g., 1965, pp. 63–67) has questioned the advisability of applying such rules. Yet the possibility cannot easily be dismissed that voluntary processes play a contaminating role.[9] The present study, like others before it, achieved some limited success in setting up rules for ridding data of the influence of V responses. But as we have seen, these rules stand upon assumptions which remain questionable and open to examination in future research. It certainly cannot be argued that further research efforts and refinements of the approach represented here will not bring improvement in the situation. Yet the thought remains that we will not be able to decide with confidence in the near future whether eyeblink conditioning can profitably be studied as a clear example of classical conditioning as understood here.

This discussion, of course, pertains to eyeblink conditioning experiments in which the UCS is an air puff; it is with such a UCS that the avoidance possibility arises.[10] Possibly the problems raised concerning voluntary processes may be circumvented by employing as a UCS an electric shock or other aversive stimulus to elicit the UCR. Several of the early studies of eyeblink conditioning did use shock (e.g., Cason, 1922). Some fairly unsystematic exploratory work in the writer's laboratory has shown rather large amounts of "adaptation" of the UCR to both shock and auditory stimuli as unconditioned stimuli. It is quite possible, however, that effective techniques can be found. We might expect that to the extent such techniques eliminate the possibility of

[9] Prokasy (1965) has recently addressed himself to some of the instrumental-like features of eyeblink conditioning.

[10] Since the present study was carried out, Spence and his associates (e.g., Spence, Homzie, & Rutledge, 1964) have reported on a "masking procedure" designed to preclude *S*'s recognizing that he is in an eyeblink conditioning experiment. This procedure has thus far been used mainly to eliminate "inhibitory sets" at the outset of extinction, but it presumably has some bearing on the presence of voluntary responding during acquisition as well.

avoidance, fewer *S*s would be discarded with the criteria developed in the present report.

It was suggested earlier that among all conditioning situations employed, eyeblink conditioning, as usually carried out, is one of the most likely to involve instrumental processes. Other workers have found other grounds for doubting the value of studying eyeblink conditioning. In a volume directed largely to the analysis of the role of classical conditioning in behavior, Mowrer (1960) devoted but one page to "short-latency" reactions such as the eyeblink. According to Mowrer, the really clear examples of classical conditioning involve "emotional" reactions. The short-latency reactions, on the other hand, are "of comparatively little biological importance and do not, apparently, show learning in its most typical important form [Mowrer, 1960, p. 386]." Razran seemed to express much the same view, listing the eyeblink among conditioned responses which are "organismically inconsequential" and "not within the center of the organism's needs and action [Razran, 1961, p. 99]."

We need not agree at once with these assessments of eyeblink conditioning. It will suffice to devote some thought and research effort to the issues raised by these views and by the analyses presented in the present report. If these considerations do not dictate abandoning eyeblink conditioning, certainly they do not suggest complacency. Whatever the alternatives, the suggestion is strong that when we conduct eyeblink conditioning experiments we do not, in an important sense, know what we are doing.

SUMMARY

Presumptive evidence exists for the presence of avoidance responding in some eyeblink conditioning experiments. Previous research led to procedures for discarding voluntary *S*s from such experiments if these *S*s exhibited large numbers of V responses defined by criteria based on response latency or response slope. The latency criterion was used in experiments which employed a ready signal; the slope criterion was used in experiments which did not employ a ready signal.

The first part of the present study was designed to determine the relation between response slope and latency in experiments using a ready signal and to determine whether the conventional slope criterion is equivalent to the conventional latency criterion in such experiments. The results showed that latency and slope criteria are not equivalent bases for identifying V responses in experiments with a ready signal.

The second part of the study was concerned with developing new criteria for identifying V responses. Essentially, the procedure involved isolating true V and C responses by instructing different *S*s to blink or not blink to the CS. Analyses of the slope and latency data of these *S*s showed that the conventional criteria were highly questionable because they placed relatively too little weight on the error of including a true V response. New criteria were specified so as to hold at .05 the estimated probability of such an error. Unfortunately, the usual procedure of discarding *S*s became highly ineffective with the new definitions of V responses. The most defensible procedure hit upon was based on a direct analysis of the median latencies and slopes of the *S*s instructed to blink or not to blink.

The eyeblink situation is far from the simple instance of classical conditioning which at first it appears to be. The value of available methods for removing the influence of voluntary processes from data obtained in this situation remains equivocal. The complications displayed in the present paper must be considered in evaluating the eyeblink experiment as a technique for the study of classical conditioning.

REFERENCES

CASON, H. The conditioned eyelid reaction. *Journal of Experimental Psychology*, 1922, 5, 153–195.

FISHBEIN, H. D. Studies in efficiency: Muscle action patterns in reaction time as related to inhibition of eyelid conditioning. Unpublished doc-

toral dissertation, University of Pennsylvania, 1963.

GOODRICH, K. P. Effect of a ready signal on the latency of voluntary responses in eyelid conditioning. *Journal of Experimental Psychology*, 1964, **67**, 497–498. (a)

GOODRICH, K. P. Invariance of eyeblink conditioning over conditions of spatial variability in a visual CS. *Psychological Reports*, 1964, **15**, 467–470. (b)

GOODRICH, K. P., MARKOWITZ, J., & NORMAN, D. A. An AC-amplification system for recording eyeblinks and other movements. *American Journal of Psychology*, 1964, **77**, 127–128.

GOODRICH, K. P., MARKOWITZ, J., & WALL, A. M. Differential eyeblink conditioning in voluntary and nonvoluntary subjects. *Psychological Reports*, 1963, **13**, 723–730.

GORMEZANO, I. Yoked comparisons of classical and instrumental conditioning of the eyelid response; and an addendum on "voluntary responders." In Wm. F. Prokasy (Ed.), *Classical conditioning: A symposium*. New York: Appleton-Century-Crofts, 1965. Pp. 48–70.

GORMEZANO, I., & MOORE, J. W. Effects of instructional set and UCS intensity on the latency, percentage, and form of the eyelid response. *Journal of Experimental Psychology*, 1962, **63**, 487–494.

GRANT, D. A., & SCHIPPER, L. M. The acquisition and extinction of conditioned eyelid responses as a function of the percentage of fixed-ratio random reinforcement. *Journal of Experimental Psychology*, 1952, **43**, 313–320.

HARTMAN, T. F., GRANT, D. A., & ROSS, L. E. An investigation of the latency of "instructed voluntary" eyelid responses. *Psychological Reports*, 1960, **7**, 305–311.

HARTMAN, T. F., & ROSS, L. E. An alternative criterion for the elimination of "voluntary" responses in eyelid conditioning. *Journal of Experimental Psychology*, 1961, **61**, 334–338.

HULL, C. L. *Principles of behavior*. New York: Appleton-Century-Crofts, 1943.

KIMBLE, G. A. *Hilgard and Marquis' conditioning and learning*. New York: Appleton-Century-Crofts, 1960.

MOWRER, O. H. *Learning theory and behavior*. New York: Wiley, 1960.

MOWRER, O. H., & AIKEN, E. G. Contiguity vs. drive-reduction in conditioned fear: Temporal variations in conditioned and unconditioned stimulus. *American Journal of Psychology*, 1954, **67**, 26–38.

MOWRER, O. H., & SOLOMON, L. N. Contiguity vs. drive-reduction in conditioned fear: The proximity and abruptness of drive-reduction. *American Journal of Psychology*, 1954, **67**, 15–25.

NORRIS, E. B., & GRANT, D. A. Eyelid conditioning as affected by verbally induced inhibitory set and counter reinforcement. *American Journal of Psychology*, 1948, **61**, 37–49.

PEAK, H. An evaluation of the concepts of reflex and voluntary action. *Psychological Review*, 1933, **40**, 71–89.

PROKASY, W. F. Classical eyelid conditioning: Experimenter operations, task demands, and response shaping. In W. F. Prokasy (Ed.), *Classical conditioning: A symposium*. New York: Appleton-Century-Crofts, 1965. Pp. 208–225.

RAZRAN, G. The observable unconscious and the inferable conscious in current Soviet psychophysiology: Interoceptive conditioning; semantic conditioning, and the orienting reflex. *Psychological Review*, 1961, **68**, 81–147.

SPENCE, K. W. Anxiety (drive) level and performance in eyelid conditioning. *Psychological Bulletin*, 1964, **61**, 129–139.

SPENCE, K. W., HOMZIE, M. J., & RUTLEDGE, E. F. Extinction of the human eyelid CR as a function of the discriminability of the change from acquisition to extinction. *Journal of Experimental Psychology*, 1964, **67**, 545–552.

SPENCE, K. W., & ROSS, L. E. A methodological study of the form and latency of eyelid responses in conditioning. *Journal of Experimental Psychology*, 1959, **58**, 376–381.

SPENCE, K. W., & TAYLOR, J. Anxiety and strength of the UCS as determiners of the amount of eyelid conditioning. *Journal of Experimental Psychology*, 1951, **42**, 183–188.

(Received May 10, 1965)

Vol. 80, No. 15 Whole No. 623, 1966
Psychological Monographs: General and Applied

PERCEPTION OF BEHAVIOR IN RECIPROCAL ROLES: THE RINGEX MODEL[1]

URIEL G. FOA
University of Illinois[2]

The ringex is a model of the cognitive organization of interpersonal behavior in reciprocal roles. A simple scheme is presented which defines 64 interpersonal variables in terms of 6 dichotomous facets. The relationship among the variables is predicted from their sequence of differentiation in the child. It is also suggested that, as the individual matures, the initial developmental pattern of interrelationship is modified by the influence of the behavior of the other person. These hypotheses appear to be supported by data referring to the reciprocal roles of husband and wife. In the discussion of the finding, an attempt is made to integrate developmental, cognitive, cross-cultural and abnormal aspects of interpersonal behavior.

The role structure of the family has been described in another paper (Foa, Triandis, & Katz, 1966), which supplied a broad picture of the relationship between these roles, but gave no details about the structure of each of them. Here only two of the family roles are considered, husband to wife and wife to husband. A closer, more detailed view of their inner structure and of their interrelationship is, however, presented, which may serve as a model for other reciprocal roles as well. Thus, while the empirical data refer to the husband and wife roles, the theoretical treatment is concerned with reciprocal roles in general.

The picture any of us has of his relationship to another person, in reciprocal roles, looks amazingly complex: it contains the behavior of the observer himself and the behavior of the other; the actual behavior and the corresponding norm; the perception and norms from the observer's point of view; and the perception and norms the observer ascribes to the other. Each of these perceptions refers to a given type of behavior. There are several types of behavior to be considered: behavior toward the self and behavior toward the other; behavior concerned with affect and behavior concerned with status; behavior which takes away and behavior which gives.

It is unlikely that the observer records these perceptions of different types of behavior in a haphazard manner. If this were the case, it would be difficult for him to compare his behavior towards the other with the other's behavior towards him, actual as opposed to ideal behavior; and so on. This type of reference is in constant use in daily life. It is probable that these different perceptions appear in a certain order and form an organized picture of our observer's relationship with the other. This study is an attempt to describe this organization.

Definition of the Variables

The picture we are going to describe is not the only possible one, nor is it the most comprehensive. Other variables could have

[1] This study has been supported, in part, by Grant M-2669 of the National Institute of Mental Health, National Institutes of Health, Public Health Service, and, in part, by a grant from the Lucius N. Littauer Foundation, New York. A preliminary version of this paper has been presented at the European Conference of Experimental Social Psychologists, Sorrento, Italy, December 1963.

Comments and suggestions on an earlier version of the manuscript, made by Aaron Antonovsky, Louis Guttman, Alex Inkeles, Elihu Katz, Uzi Peled, and Harry Triandis are gratefully acknowledged. So is Colin Kessel's help in editing the manuscript.

[2] On leave from the Israel Institute of Applied Social Research.

been chosen and more added to those defined. The selection of our variables, analyzed in this study, can be justified on three counts:

a. They are based on notions which are well established in sociopsychological literature;

b. They are generated in a systematic manner rather than chosen arbitrarily;

c. They lead to the prediction of empirical results.

Each of the variables discussed here defines a certain perception of a certain behavior. Indeed, each type of behavior is perceived in several different ways and each type of perception may refer to different behaviors. The variables can therefore be cross-classified according to the type of perception and the type of behavior. It follows that in order to define the variables it will be sufficient to define types of perception and types of behavior. The combination of any of these two types will produce a variable.

The Perceptual Types

Consider any interpersonal behavior, like, for example, friendliness to the other. One does differentiate between his friendliness to the other and the friendliness of the other to him; between actual friendliness and the corresponding norm; between friendliness as perceived from his point of view, or according to his norm; and friendliness according to the point of view and norm ascribed to the other.

More formally, the perception of a given behavior is differentiated according to the following three perceptual facets:

A. The person doing the action, or *actor*, with two elements: a_1, the other (nonobserver), and a_2, the self (observer);

B. The *level:* b_1, actual (what is done), and b_2, ideal (what ought to be done);

C. The person from the point of view of whom the action of a given actor is perceived, or alias: c_1, the other (nonactor), and c_2, the self (actor).

Thus, a given perception of our observer can be classified according to the actor, the level, and the alias.

The profiles of the elements of the facets taking one element from each facet define eight perceptual types. For example: $a_1b_1c_1$ is the perception of the actual behavior of the other from the point of view of the other; $a_2b_2c_1$ is the perception of the ideal behavior of self from the point of view of the other, and so on.

To call "what ought to be done" a perception is, perhaps, to stretch the meaning of this word. It would have been more appropriate to use a term like norm, value, ideal. We have used the word perception in this sense for lack of a better term covering both actual and ideal behavior.

The Behavioral Types

The types of behavior are defined, in a manner similar to the perceptual types, by the following three facets:

D. *Content* of behavior: d_1, acceptance or giving, and d_2, rejection or taking away.

E. *Object* of behavior: e_1, the other (nonactor), and e_2, the self (actor).

F. *Mode* of behavior: f_1, social or status, and f_2, emotion or love.

Taking profiles over the elements of these facets eight behavioral types are defined, as for example, $d_1e_2f_2$ or emotional acceptance of self; $d_2e_1f_1$ or social rejection of other.

This classification of behavior suggests the giving or taking away of love and status from the self and the other as basic features of interpersonal interaction. Other conceptions of this behavior are of course possible and have been proposed. The reasons for choosing this particular formulation, on the basis of theoretical considerations and empirical findings, have been discussed at length in an earlier paper (Foa, 1961). Additional findings supporting this formulation have been reported by Adams (1964).

Combining the Types

By combining any perceptual type with any behavioral type we can define a variable. For example: a combination of perceptual type $a_2b_1c_2$, the perception of the actual behavior of the self from the point of view of the self; with behavioral type $d_1e_1f_2$, emotional acceptance of the other, produces $a_2b_1c_2d_1e_1f_2$, which is the degree

(intensity, frequency) of our observer's perception of his actual emotional acceptance of the other. The notion of degree or intensity or frequency, although not explicit in the formal notation, has been added to indicate the range of values that a variable can assume. In the same manner other variables can be defined, just by combining a perceptual type with a behavioral one. Since there are eight of each type, $8 \times 8 = 64$ variables have been defined (see Table 1).

Each row of the table defines a perceptual type, indicated by a Roman numeral from I to VIII. Each column of the table defines a behavioral type, numbered from 1 to 8. Each cell of the table, at the crossing of a given row with a given column, defines a variable made up of the perceptual type of this row and of the behavioral type of this column. To indicate a variable is, therefore, sufficient to specify the numerals of its perceptual and behavioral types. For example, Row I, Column 1 defines the Variable I, 1 "To what degree does the observer perceive, from his point of view, that the other (nonobserver) accepts him socially," or "To what extent the observer perceives he is receiving status from the other."

Some other examples are: Variable V, 2 "To what degree, according to the observer, does the other think that he (the observer) ought to accept the other emotionally."

Variable VII, 5: "To what degree does the observer feel that he is rejecting himself socially (taking away status from himself)."

Variable VI, 7: "To what degree does the observer feel that he ought to reject the other emotionally (to deny him love)."

The upper half of the table refers to the role of the other toward the observer, the lower half to the role of the observer toward the other.

This classification of variables can easily be interpreted in terms of social exchange. Variables in Rows I–IV, Columns 1–2 and 7–8, refer to what the nonobserver gives to the observer, that is, to what the observer receives from the other. Variables in Rows V–VIII, same columns, concern what the observer gives to the other. The remaining variables refer to what each actor gives to himself. Two currencies are used, love and status. What is and what ought to be given and denied is recorded from the point of view of the observer as well as from the point of view of the other.

In the eight variables in any given column the behavioral type is constant, only the perceptual type changes. In the variables of the same row, on the other hand, the perception is constant, while only the

TABLE 1

FACET DEFINITION OF THE VARIABLES OF AN OBSERVER

Perceptual Type				Behavioral Type							
			Content	Acceptance				Rejection			
			Object	Other		Self		Self		Other	
Actor	Level	Alias	Mode	Social	Emotional	Emotional	Social	Social	Emotional	Emotional	Social
			Type	1	2	3	4	5	6	7	8
Nonobserver or other	Actual	Nonactor	I								
		Actor	II								
	Ideal	Actor	III								
		Nonactor	IV								
Observer or self	Ideal	Nonactor	V								
		Actor	VI								
	Actual	Actor	VII								
		Nonactor	VIII								

behavior changes. Thus, for example, the variables of Row VII refer to the degree to which each one of the eight behavioral types appears in the observer's perception of his own actual behavior from his point of view.

We can, therefore, look at the 64 variables column by column keeping the behavioral type constant or row by row holding constant the perceptual type. This manner of classification will prove useful in the formulation of hypotheses about the interrelationship among variables.

The Hypotheses

Each one of the 64 variables may have a certain frequency or intensity, varying from very little to very much, or from rarely to often. The hypotheses are concerned with the relationship among the frequencies of the variables. It is proposed that the eight variables belonging to any given row or column of Table 1 will be related in the same order in which they appear in the table; that is, the nearer any two variables are in the table the higher will be their correlation. It is further suggested that the first and last variable of a given row or column will also be fairly close to each other. So the intercorrelation matrix of each set of eight variables will approximate the circumplex pattern (Guttman, 1954). To explain how the hypothesis was derived some comments on the sequence of development of notions of interpersonal perception in children appear appropriate.

Let us start with a very simple consideration. When the adult observer of our study interacts with another person he is apparently able to classify the ongoing behavior in the scheme that we have presented. He is able to differentiate between what he does and what the other fellow does, between what is being done and what ought to be done, between acceptance and rejection, and so on. On the other hand, when a newborn child becomes an observer he can make none of these differentiations. What is then the sequence of development which bridges this enormous gap? Which concepts develop first and which next? We shall make some proposals regarding the sequence of development of the perceptual and behavioral facets which have been defined.

It is proposed that, in the perceptual facets, the differentiation between actors develops first, followed by the differentiation between levels and then between aliases. The differentiation between self and nonself, as actors, seems to occur, in the pre-oedipal phase, as soon as the child realizes that his behavior and the behavior of his mother are not one and the same thing. In fact this differentiation provides the child with his first two roles: his role toward his mother and the role of the mother toward him. The differentiation between actual and ideal level could not be easily made before the actors differentiation: ideal behavior, in its elementary form, is the behavior of the child which is followed by acceptance behavior of the adult; thus it seems to require differentiation between the two actors, the child and the adult. The realization that "what mother wants me to do" may be different from "what I want to do," provides the beginning of the third differentiation, between the point of view of the actor and the point of view of the other. Therefore, if differentiation by level requires prior differentiation by actor and contains the beginning of differentiation by alias, the suggested sequence of development will be sustained: Nonself is differentiated from self, as actor, then ideal behavior from actual behavior and, finally, the point of view of the other from the point of view of the self. This suggests that initially the child perceives that he is the actor of everything being done. Self, as actor and alias, and actual are the primary elements of the perceptual facets. These elements, however, become meaningful only after their differentiation from the other secondary elements. As Piaget (1955, p. 237) has noted "Precisely because he feels omnipotent, the child cannot yet contrast his own self with the external world."

Among the behavioral facets the proposed sequence is: first, content; then, object; and finally, mode. There are a number of considerations suggesting that the differentiation between accepting and re-

jecting may be made very early in child life. It may well be the very first interpersonal differentiation made by the child. Omologous physiological mechanisms, like inhaling and exhaling, suckling and excreting, are present from birth. In early psychomotoric development the child first learns to grasp, then to reject things, thus establishing a distinction between these two types of action. The same sequence occurs in the first days of life of imprinting animals. Furthermore, it seems that some elementary code for classifying and recording behavior, like accepting and rejecting, is necessary before the notions of object and actor become meaningful. The notion of *what* is done seems likely to precede the notion of *to whom* is done (object) and *by whom* (actor). On the other hand the differentiation of status from affect requires the perception of a social group larger than the mother-son dyad or the mother-father-son triad, in order to occur. Even if the child realizes that the parents have higher status than himself, this status difference will be overlapping with the self-other differentiation. It is only when the siblings enter the social scene of the child that the notion of status may have a chance to become established: there are now others who have higher status than the child (parents) or a similar one (siblings). These considerations suggest that differentiation of rejection from acceptance will occur first and differentiation of status from affect last, with the

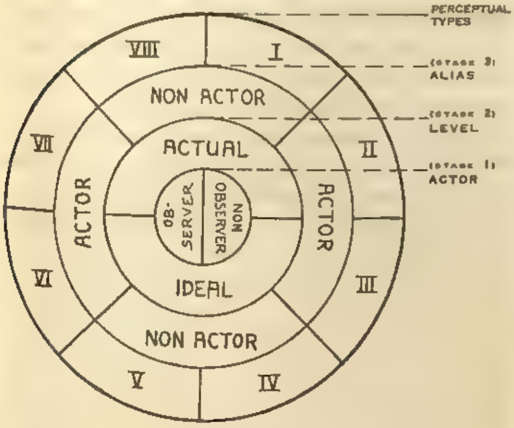

FIG. 1. Assumed stages of differentiation of perceptual types.

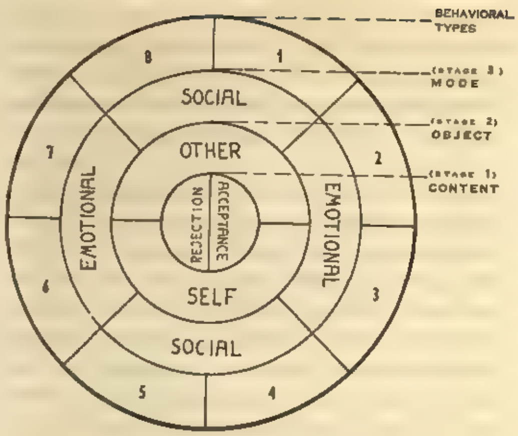

FIG. 2. Assumed stages of differentiation of behavioral types.

differentiation of other from self, as objects, occurring at some intermediate time.

These three stages of this proposed process of concept building by binary fission are schematically represented in Figure 1 for the perceptual facets and in Figure 2 for the behavioral facets.

In Figure 1 the inner circle indicates the differentiation by actor (observer and non-observer). Moving away from the center, the next circle shows how the differentiation by actor subdivides again according to the level: actual and ideal. The third circle indicates the further subdivision by alias. The outer circle shows the eight perceptual types numbered from I to VIII resulting from the three stage process of successive subdivision of concepts. Each successive differentiation is obtained by a subdivision of the previous concept according to the new facet.

An identical process is shown in Figure 2 with regard to the behavioral facets. The inner circle indicates the differentiation by content between acceptance and rejection. The next one between objects and the third one between modes. The outer circle shows the eight behavioral types, numbered from 1 to 8, resulting from this process.

The concept of observer is not included in this developmental scheme. A child learns to differentiate between actors, aliases, objects, modes, etc. To do this he has first to *become* an observer, but no differentiation

between observers occurs at any stage of development. What in common parlance is often described as differentiation between observers is in fact differentiation between points of view of the same observer.

The perceptual types of Figure 1 and the behavior types of Figure 2 follow the same order in which they are given in Table 1. The figures, show, however, that the proposed order of types seems to result from the process of successive differentiation of concepts occurring during childhood. Having thus explained how the hypothesis was derived, we can now turn to the procedure for gathering the data and to the presentation of the findings.

PROCEDURE

The empirical evidence to be presented is provided by a study of a sample of 633 married couples in Jerusalem, Israel. Husband and wife were interviewed separately and simultaneously in their home by two field workers.

Each one of the 64 variables defined in Table 1 was observed by means of a three-question Guttman scale. The scale score indicates the degree of perceived occurrence of the variable, as reported by the observer. A score was therefore obtained for each respondent on each one of the variables. In fact, the questionnaire was built according to the facet design of the 64 variables as given in Table 1. Let us explain how this was done: For each one of the eight types of behavior and for each actor, three brief stories were prepared. For example, the three stories referring to the husband's social acceptance of the wife run as follows:

Abraham has consideration for his wife and displays toward her respect and esteem.

Isaac thinks his wife is very successful and especially esteems her personality and her actions.

Jacob is sure that everything his wife does is important and good and there is no limit to the esteem and importance that he attributes to her.

Some other examples of stories are:

Social acceptance of self: Abraham is a husband who esteems himself and relies on himself and on his decisions.

Emotional acceptance of self: Isaac is a husband who is satisfied with his actions and feels very much at peace with himself.

Social rejection of wife: Abraham slightly criticizes his wife's behavior and thinks that she makes a few mistakes.

Emotional rejection of self: Jacob is a husband very dissatisfied with himself and with his behavior toward his wife, rejects and blames himself.

Similar stories were used for the behavior of the wife.

After each story four questions were asked, to differentiate between perceptual types, as follows:

Actual level, alias of the actor

Do you behave toward your wife as does the husband in the story? (Almost always; generally; sometimes; seldom; almost never.)

Ideal level, alias of the actor

Do you think that a husband should behave as does the husband in the story in relation to his wife?

Actual level, alias of the nonactor

Would your wife say that you resemble the husband in the story?

Ideal level, alias of the nonactor

Would your wife say that a husband should behave thusly?

In this manner, the facet definition of the variables (see Table 1) provided the basis for constructing the questionnaire: the behavioral type and the actor were specified by the story, and the remaining two facets of the perceptual type, level and alias, by the question. The same question was asked three times: once after every one of the three stories for a given type of behavior and actor. Thus three observations were obtained for each variable. They were found to be scalable, and a scale score was computed following the usual procedure for Guttman scaling. The scale score of each variable was then correlated with the score of other variables.

THE EMPIRICAL STRUCTURE

The intercorrelations between the eight variables belonging to the same row or column of Table 1 can be arranged in a 8 × 8 matrix. Since there are, in Table 1, eight rows (and eight columns), it follows that eight intercorrelation matrices are obtained for each observer when the variables are taken either by row or by column. The study includes two observers, husband and wife. Thus, there are, in total, 16 matrices to be considered for the rows and an equal number for the columns.

It has been predicted that the intercorrelations between the eight variables, in the same row or in the same column of Table 1, will tend to follow the circumplex pattern. In a circumplex the higher correlations are found near the main diagonal; moving away from the diagonal cell the coefficients decrease and then increase again. As already noted, the circumplex suggests a circular order of the variables, so that the first and last variables are also closely related (Guttman, 1954). When the order is open, so that

TABLE 2

AN EXAMPLE OF BEHAVIORAL CIRCUMPLEX; CONSTANT PERCEPTUAL TYPE VII; OBSERVER: WIFE

Type	1	2	3	4	5	6	7	8
1	—	65	27	24	06	20	35	45
2	65	—	28	16	09	20	36	40
3	27	28	—	52	28	17	01	07
4	24	16	52	—	31	17	−11	−01
5	06	09	28	31	—	39	18	24
6	20	20	17	17	39	—	34	35
7	35	36	01	−11	18	34	—	53
8	45	40	07	−01	24	35	53	—

Note.—Decimal point omitted in all tables.

the first and last variables correlate least, the pattern is called simplex (Guttman, 1954). Common to these two intercorrelation patterns is the notion of order among the variables (Guttman, 1958).

The Order of Behavior Types

Let us consider the intercorrelations among variables belonging to the same row of Table 1. In each one of these sets of eight variables the perceptual type and the observer are constant, while the behavioral type changes. Thus, each matrix shows the relationship among the eight types of behavior for a given perceptual type, that is, for a given actor, level and alias of one of the two observers.

The example of Table 2 refers to perceptual Type VII of the observer wife. The full set of 16 matrices has been published earlier (Foa, 1962) and is also reported in the appendix.[3] To give the variable scores a single meaning, from unfavorable (low acceptance or high rejection) to favorable (high acceptance or low rejection) to the interpersonal relation, the scores of the rejection variables were reversed. This explains the positive correlation coefficient between acceptance and rejection: it just means that the more frequent the acceptance the less frequent the rejection.

The predicted order of behavioral types is rather well supported by the correlations of Table 2: the coefficients are high near the main diagonal and then decrease and increase again. In the first row of the table, for example, the coefficients decrease as one moves from the first column to the right, reach the lowest point in Column 5, then increase again gradually. In this matrix there are four deviations from the predicted pattern: the coefficients between Variables 1-7, 2-4, 5-7, and 6-7 are lower than expected. Some of the other 15 matrices are even better than this one, having as few as one or two deviations. The largest number of deviations is nine (in two cases), and the median number of deviations for all the 16 matrices is six. Some of these deviations tend, however, to be systematic: correlation between Type 2 and Type 4, for example, is always lower than expected. The correlation between Type 6 and Type 7 also is often too low.

It will be noted that contiguous variables are not equally spaced: types of behavior referring to the same object (self or other) like 1–2 and 3–4, are usually nearer or more correlated than behavior types referring to a different object, as 2–3 and 6–7. The correlation between social acceptance and social rejection of self (Types 4 and 5) is also often low. These features are, however, quite compatible with the predicted order.

This order suggests certain predictions with regard to the relationship between acceptance and rejection, as well as between self and other, which may be of interest. Ambivalence of emotional feeling (accept-

[3] See the technical appendix to this paper which has been deposited with the American Documentation Institute. Order Document No. 9026 from ADI Auxiliary Publications Project, Photoduplication Service, Library of Congress, Washington, D. C. 20540. Remit in advance $2.50 for microfilm or $6.25 for photocopies, and make checks payable to: Chief, Photoduplication Service, Library of Congress.

ance and rejection) occurs at an early stage of development of the child, before the social mode is differentiated from the emotional one. The order of behavior types suggests that, even in the adult, more ambivalence is possible at the emotional mode than at the social one: one may give and deny affect to the same person, but status requires a more clear-cut decision. In the proposed order social acceptance and rejection of other (Types 1 and 8) and of self (Types 4 and 5) are neighbors and therefore expected to be closely (and inversely) related. On the other hand, emotional acceptance and rejection of other and self (Types 2–7 and 3–6) are farther apart in the order and probably less interrelated than the corresponding social types. Inspection of the 16 intercorrelation matrices shows that this prediction is supported in 26 out of 32 comparisons. Interestingly enough, all the six deviations, except one, occur in the matrices referring to the point of view of the other, that is the point of view which develops *later* in the child.

The proposed order of behavior types also suggests that self and other will be more related at the emotional mode than at the social one. Types 2–3 and 6–7, concerning emotional behavior toward self and other, are indeed neighbors, while the corresponding social types, 1–4 and 5–8, are not. This hypothesis, like the previous one, rests on a developmental rationale: the differentiation between self and other, like the differentiation between acceptance and rejection, is less strong at the primary emotional mode than at the social mode, which appears later in the development sequence. Inspection of the 16 matrices shows that this hypothesis is supported only in half of the cases, the result expected under chance conditions. When, however, the matrices are classified according to the facets of their constant perceptual type some interesting differences emerge. In the matrices referring to actual behavior of self (an early type of perception), the relationship between self and other is always higher at the emotional than at the social level, as predicted. The hypothesis is not supported in the matrices referring to the later perceptions (ideal behavior and behavior of other).

The matrix of Table 2 refers to the perception of the actual behavior of self from the point of view of self. In terms of development, this is perhaps the first type of perception that occurs to a child. It is not without significance that in this matrix, which refers to the observer wife, as well as in the corresponding matrix for the husband, both the above hypotheses are fully supported. Furthermore, in the corresponding matrix for the husband, there is only one deviation from the circumplex pattern. Thus, while the predicted order of behavior types seems to be supported by the 16 matrices, the tendency to deviate from it may be somewhat stronger in matrices referring to types of perception which occur later in the development of the child. A closer analysis of these deviations suggests tentatively that they may be due to a tendency of the facets to become less interdependent, to break away from the hierarchical pattern resulting from the development sequence. Thus the structure would move toward the cubex (Foa, 1965) each one of the three facets acquiring a dimension of its own. If this is true, it may be possible that better circumplexes will be found in younger individuals, with increasing deviations toward the cubex model, as the individual becomes more mature. Maturity, in this sense, would mean that the criteria of behavioral differentiation become more independent from their sequence of development in childhood.

The Order of Perceptual Types

In the variables appearing in the same column of Table 1, the behavior type and observer are constant, while the perceptual type changes. Thus the matrix of intercorrelations of these variables shows how the frequency of the given behavior in one actor is related to the frequency of the same behavior in the other actor, both at the actual and ideal levels, and from the points of view of both aliases. Again there are 16 matrices to be considered, one for each column of Table 1 for the observer husband and the same number for the observer wife

(See Footnote 3). One example is given in Table 3. It refers to behavioral Type 2, emotional acceptance of the other, of the wife. In this example, the predicted order of perceptual types is very well supported. There is only one deviation, the correlation between III and VI being lower than expected. In all the other matrices the number of deviations is higher: as much as 10 in one case and a median of 5 for all the matrices.

In all these matrices the relationship between actual and ideal behavior is always higher for the point of view of the actor (Types II–III and VI–VII) than for the point of view of the other (Types I–IV and V–VIII). The developmental explanation, suggested for the behavior types, may serve again here: actual-ideal differentiation is less sharp for the primary element of the alias facet, the actor, than for the element, other, which appears later.

These perceptual matrices also show that the relationship between the behaviors of the two actors is closer at the actual level than at the ideal level when the object of the behavior is the other, that is, in the matrices referring to the two first and last columns of Table 1. When the behavior is directed toward the self (Columns 3 and 6 of the same table), the contrary happens: the behaviors of the two actors are related more at the ideal level than at the actual one. The correlation between Types I and VIII (actual level) is higher than the correlation between IV and V (ideal level) when the object of behavior is the other, and lower when the object is self. This occurs in all the matrices, except the one of Table 3. Thus the relationship between the reciprocal behaviors of the two actors is stronger than the relationship between the respective norms. When, however, the behavior is directed toward oneself, the norms of the two actors are related more than their actual behaviors. In the latter case, Types I and VIII are rather apart, so that the order looks more like a simplex than a circumplex. In the former case, however, the distance between Types IV and V is not large enough to change the tendency toward a circular order.

In spite of the gaps just noted, these perceptual matrices indicate a fairly close relationship between the behaviors of the two actors, husband and wife. Elsewhere (Foa, Triandis, & Katz, in press), it has been shown that such relatively high relationship can be expected when the two actors occupy similar positions in the power structure of the family system. This study also suggests that the correlations between the two actors are likely to be lower when there is more difference in power between the two roles, as for example, between father and son, and possibly more so for the behavior types dealing with status than for those dealing with affect.

So far we have considered the relationship among the same perception of different types of behavior and the relationship among different perceptions of the same type of behavior. The results tend to support the hypotheses of order among these types. An attempt has been made to explain certain features of this order as well as some fairly systematic deviations from it, according to the same rationale which served for generating the hypotheses: the sequence of

TABLE 3

AN EXAMPLE OF PERCEPTUAL CIRCUMPLEX; CONSTANT BEHAVIORAL TYPE II; OBSERVER: WIFE

Type	I	II	III	IV	V	VI	VII	VIII
I	—	71	59	42	00	26	45	52
II	71	—	67	55	14	30	40	51
III	59	67	—	61	26	34	31	44
IV	42	55	61	—	53	46	32	36
V	00	14	26	53	—	47	33	33
VI	26	30	34	46	47	—	56	47
VII	45	40	31	32	33	56	—	61
VIII	52	51	44	36	33	47	61	—

development of interpersonal perception in the child. These features of the order of types will have to be taken into account as we turn to the next problem, the relationship among variables when both the perceptual and behavioral facets are different, that is, variables belonging to different rows and columns of Table 1. Thus, for example, relating what we receive from the other to what we give to ourselves, involves variables in different rows and columns.

Central and Peripheral Variables

The prediction of the relationship between variables taken from different rows and columns of Table 1 seems to require some understanding of the manner in which the set of 64 variables is organized as a whole. For this purpose, it may be useful to consider the position of each variable with respect to the set, as indicated, for example, by the multiple correlation of a variable to the other ones. Some of these multiple correlations may be higher and some may be lower. Essentially this is the problem of communality in factor analysis (Foa, 1963). It has been recognized that the communality of a variable is relative to the set of variables to which it belongs rather than a property of the variable itself. The set of variables considered in this study is interpersonal behavior. Thus all the variables are interpersonal, but, to paraphrase George Orwell, some variables may be more interpersonal than other ones. If it is so, we may expect that the multiple correlation of the more interpersonal variables will be higher: they belong to the set more than the other, less interpersonal ones.

To decide the degree to which a variable is interpersonal consider its facet elements. Each facet has two elements. These pairs of elements are: self-other; emotional-social; acceptance-rejection; and actual-ideal. In each pair one of the two elements appears to be more interpersonal than the other one. Relations with *other* are characteristic of interpersonal behavior, while relations with self are not. *Social* behavior is interpersonal, while emotional behavior need not be so. Some degree of *acceptance* is necessary for the interpersonal relationship to continue, while rejection leads to the cessation of the relationship. There is no interpersonal situation without *actual* behavior, but an ideal can be maintained without reference to the other. Thus, the elements social, other, acceptance, and actual, are more closely associated with interpersonal behavior than the elements emotional, self, rejection, and ideal.

The variable containing only interpersonal elements is Variable I, 1, situated at the upper left corner of Table 1: actual social acceptance of the other by the other from the viewpoint of the other. The variable containing none of these elements is VI,6: ideal emotional rejection of self by self from the viewpoint of self. It is predicted that multiple correlation will be high for the most interpersonal variable and will decrease as one moves toward the least interpersonal one. To test this hypothesis the multiple correlation of each variable with the other variables of the same behavioral

TABLE 4

Multiple Correlations of the 64 Variables for the Observer Wife

Perceptual type	Behavioral type							
	1	2	3	4	5	6	7	8
I	78	78	65	70	52	45	75	74
II	77	75	73	72	49	47	63	61
III	79	78	66	68	47	44	61	64
IV	71	67	66	65	46	46	51	56
V	68	63	56	55	51	48	47	52
VI	70	63	56	58	49	44	47	51
VII	71	68	57	60	52	51	61	64
VIII	80	77	72	68	49	56	63	67

Note.—Each variable is defined by a given profile of perceptual and behavioral type.

TABLE 5
MULTIPLE CORRELATIONS OF THE 64 VARIABLES FOR THE OBSERVER HUSBAND

Perceptual type	Behavioral type							
	1	2	3	4	5	6	7	8
I	78	74	63	66	51	45	72	75
II	78	76	70	68	52	39	67	68
III	76	71	68	66	52	46	66	64
IV	67	63	67	63	46	33	58	59
V	67	67	65	62	63	60	64	62
VI	67	63	65	67	59	59	50	52
VII	73	70	67	65	49	45	64	69
VIII	82	77	66	64	52	53	68	73

circumplex was computed. It would have been more appropriate to use all the remaining 63 variables of the set in computing each multiple correlation, but the labor involved would have been prohibitive. These multiple correlations are given in Table 4 for the observer wife and in Table 5 for the observer husband.

In Tables 4 and 5, the variables are exactly in the same order as in Table 1, but now we can interpret this order as going from the most to the least interpersonal variable, according to the component-like behavior of the facet elements (Foa, 1965). As predicted, the coefficient of multiple correlations follows this order closely: they

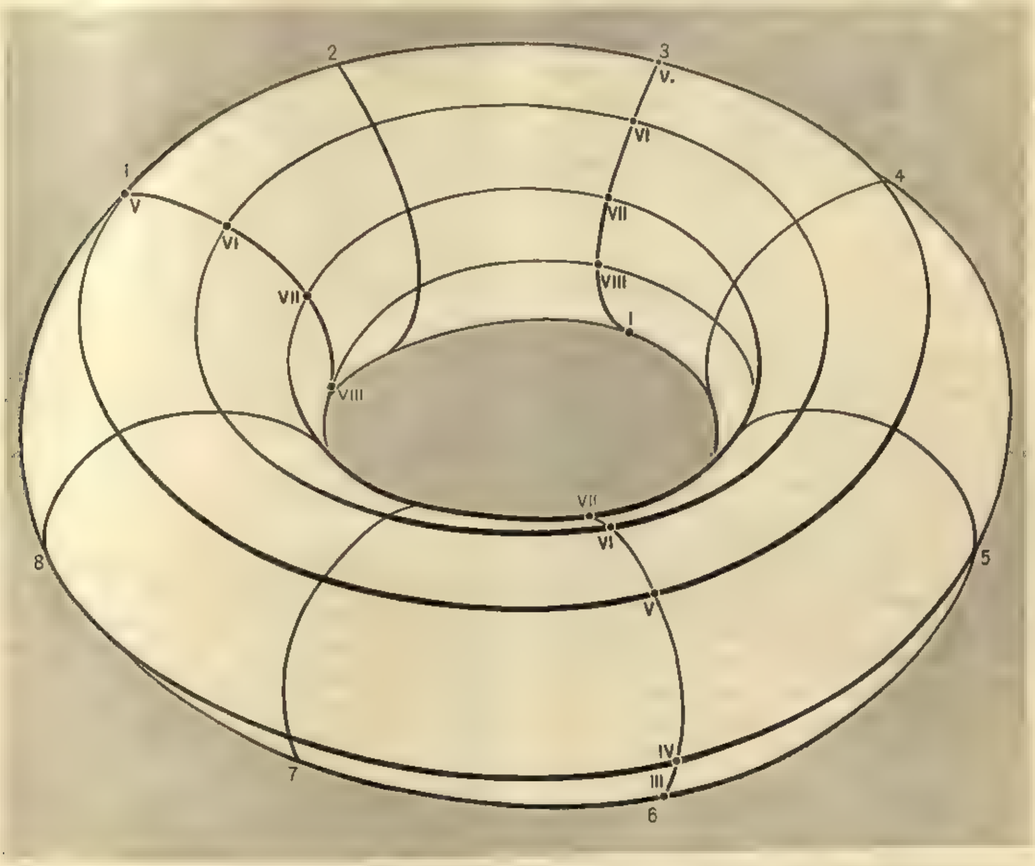

FIG. 3. A representation of the ringex model.

are high in the top left corner and decrease moving to the right or down until the sixth column or row is reached, then start increasing again. Along the diagonals of each table this trend is even more regular. Since the more interpersonal variables have a multiple correlation higher than the less interpersonal one, their position in the configuration should be more central than the position of variables which are less specifically interpersonal.

In this manner we have gained some information with regard to the configuration of the variables: it should be shaped in such a way as to permit a distinction between the center and the periphery. It should also account for the intersecting circles of eight variables which have been described previously. The anchor ring or torus, a figure shaped like the inner tube of a car, a doughnut, or a bagel, fulfills these requirements. In an anchor ring the points situated in the inner part of the surface, where the air valve is found in a car tube, are more central than those situated in the outer part. We have called a statistical structure of variables, arranged on the surface of an anchor ring, a *ringex*. The proposed ringex structure of our 64 variables is portrayed in Figure 3.

Correspondence between Conceptual and Empirical Structures

This figure is an attempt to represent the empirical relationship among the variables of Table 1. The large eight circles of the figure are the behavioral circumplexes and correspond to the rows of Table 1: each of these circles is made up of the same variables which are found in a given row of Table 1. The small circles represent the order of perceptions, each one of them corresponds to a given column of Table 1. Each behavioral circle crosses each perceptual circle once, and this point of intersection defines the position of a given variable with respect to the other ones. It corresponds indeed to a cell of Table 1. Thus, Figure 3 depicts the proposed empirical interrelationship among the variables while Table 1 depicts their conceptual relationship in terms of facet elements. It is remarkable how closely these two representations correspond to one another. In effect the ringex structure can be obtained by folding Table 1 so that the bottom row will join with the top row and then folding it again in the other dimension so that the first and last columns will also become neighbors. Reversing the order of these two operations would have produced a different ringex, but data to be presented later tend to support the ringex of Figure 3.

Evidence for the Ringex Hypothesis

If the ringex of Figure 3 is a correct portrayal of the interrelationship among the variables, it will also be expected that:

1. The average correlation among variables belonging to the same behavior circle will be highest for the circles in the internal portion of the surface and will decrease gradually as one moves toward the external part. Internal circles are indeed smaller than the external ones. The smaller the circle the larger the correlation, on the average, among its eight variables. Correlation is inversely related to distance.

2. The average correlation among variables belonging to the same perceptual circle will always be higher than the average correlation among the variables of any given behavioral circle. Perceptual circles are, in the figure, always smaller than even the smallest behavioral circle so that their variables should correlate higher.

The average correlations for each circle and for each observer are given in Table 6.

For simplicity's sake the circumplexes of Table 6 are all indicated by an Arabic numeral. It would have been more correct to denote a behavioral circumplex by the Roman numeral of the perceptual type which is constant in it. Both hypotheses appear well supported. The average correlations for the behavioral circumplexes, given in the first two columns of Table 6, are larger for the smaller circles found in the inner portion of the ringex (a large correlation indicates a small circle) and

decrease in size as one moves to the outer portion where circles are larger. On the other hand, the changes in average correlation of the perceptual circumplexes, last two columns of the Table, do not appear to be related to their order position. In fact, there is nothing in the ringex structure to suggest such a relationship.

The lowest correlation of the perceptual circles is .44; the highest one of the behavioral ones, .31. Since a large correlation indicates a small circle, this finding supports the suggestion of Figure 3 that the perceptual circles are always smaller than the smallest behavioral one.

Different perceptions of the same behavior are, on the average, more interrelated than the same perception of different behaviors. This may not be surprising considering that perceptual types relate actual and ideal behavior from the point of view of self and other. Realizing that the perception or norm of the other is different from our own, or that there is a discrepancy between actual behavior and the corresponding norm may generate stress and set in motion mechanisms for reducing the gap. As a result, correlations between actual and ideal level, as well as between points of view, tend to be relatively high. Whether or not the correlations between actors will also be high, may depend on the relative power position of the two roles and on the culture. In the same culture, father and son, who have different power positions, are less related than husband and wife, but both these pairs of roles are more related in the American culture than in the Japanese one (Foa, Triandis, & Katz, in press).

Differences among behaviors, on the other hand, may be less conducive to strain. Perceiving a discrepancy between status and love, between behavior toward self and other, and even between acceptance and rejection, may not necessarily produce strain. In fact, these differentiations appear to be embedded in the cultural values, so that members of a certain culture are trained to make those differentiations which appear to be of particular significance to their specific culture (Foa, 1964). Thus, differentiating among behaviors in a manner consonant to the cultural values may even facilitate the adjustment of the individual to this culture. This may explain why the perceptual correlations are larger than the behavioral ones.

TABLE 6

Average Correlations of the Behavioral and Perceptual Circumplexes for each Observer (Wife and Husband)

Circumplex	Behavioral		Perceptual	
	Wife	Husband	Wife	Husband
1	27	29	54	60
2	25	28	44	53
3	23	25	64	70
4	21	24	59	65
5	18	21	48	58
6	18	23	54	58
7	27	26	57	65
8	28	31	47	58

Another difference revealed by Table 6 is that the correlations for the husband are always higher than the corresponding correlations of the wife, except in one case. Comparing the multiple correlations of husband and wife, of Tables 4 and 5, leads to the same conclusion: the wife has a finer, more differentiated picture of the relationship than the husband. It may be that the wife has fewer roles in society than the husband, so she can afford to "specialize" in the particular task of wife. The husband's specialization may show itself in a greater differentiation among more different roles compensated by the avoidance of overspecialization in the specific role of husband. This explanation is consistent with certain other findings (Foa, 1964; Foa & Chemers, 1966), suggesting that there are limits to the ability of a person to differentiate between and within roles while maintaining self-identity. Thus, overdifferentiation in one area may be counterbalanced by underdifferentiation in another area.

The Relationship among Behavior Circles

The evidence presented so far provides some measure of support for the ringex model. Pursuing the empirical test further we attempt to predict the relationship be-

tween variables belonging to different behavioral or perceptual circles.

Consider the correlations among variables belonging to any two behavioral circles: they show how the eight behavior types interrelate when the actor, level and/or alias change. Some of these correlations may be of particular interest. Criteria for judging whether a particular set of intercorrelations is interesting or not may vary, but one is likely to be more interested in the relationship between behavior types which differ in one perceptual facet only, rather than in those differing in two or three facet elements. Thus, for example, the relationship between the actual behaviors of the two actors, or between actual and ideal behavior of the same actor, may be thought of as more interesting than, say, the relationship between the actual behavior of one actor from his point of view and the ideal behavior of the other actor from the point of view of the other. The advantage of the model is that all these relationships are considered in a systematic manner.

In Figure 3 the behavioral circles are shown as being roughly concentric. A variable in one circle appears to be nearest to the corresponding variable in the other circle. For example, among the variables of Circle VI, the one which appears nearest to Variable 1 of Circle V is precisely Variable 1. If this representation is correct, any variable on a behavioral circle will correlate highest with the corresponding variable on another behavioral circle; its correlation with the other variables of the second circle will then decrease and increase again, following the usual circumplex model.

Several examples of intercorrelation matrices for various pairs of behavioral circles are given in the appendix (see Footnote 3). They follow fairly closely the predicted pattern. The number of deviations is fairly small but tends to be larger when the behavior of the two actors is interrelated, as in Matrices IV–V, I–VIII, I–VI, and I–VII. One example of the intercorrelation between actors is given in Table 7. This table relates the actual behaviors of husband and wife, as perceived by the husband from his point of view. In this particular example, there is a rather large number of deviations, more than in most other matrices of this type; nevertheless the tendency toward an ordered pattern is readily apparent. This order suggests that what the husband gives (or denies) to his wife (in affect and status) is more related to what she gives (or denies) to him, than to what she gives to herself. On the other hand, what the husband gives to himself is more related to what the wife gives to herself than to what she gives to him. Within a given level, actual or ideal, behaviors toward the other go together, as do behaviors toward the selves of the two actors. The relationship between self and other becomes different, however, when the two levels are compared as done in the cor-

TABLE 7

AN EXAMPLE OF INTERCORRELATIONS BETWEEN TWO BEHAVIORAL CIRCLES

	Behavioral Type	Perceptual Type I							
		Behavioral Type							
		1	2	3	4	5	6	7	8
Perceptual Type VII	1	60	55	39	29	23	13	29	31
	2	52	53	31	25	10	15	18	26
	3	41	30	56	35	12	06	07	10
	4	38	22	39	68	08	01	03	09
	5	23	05	14	08	35	22	22	22
	6	22	21	−07	−04	23	37	29	34
	7	31	34	14	−01	31	29	57	61
	8	42	45	18	05	30	26	48	55

relations between perceptual circles, discussed in the next section.

Relationship among Perceptual Circles

Let us now turn to the relationship among variables appearing on two different perceptual circles, that is, the relationship between the various perceptions of two given types of behavior of the two actors. Some of these correlations relate the behavior of one actor toward himself to his behavior toward the other. The intercorrelations between the variables of perceptual Circles 2 and 3 indicate, for example, how emotional acceptance of self is related to feeling emotionally accepted by the other; Circles 1 and 4, on the other hand, relate giving status to self to receiving status from the other.

To develop hypotheses about the relationship among perceptual circles, Figure 3 will be used again. Consider two circles which are near to each other, Circles 1 and 2, for example. From the analysis of the behavioral circumplexes we know that these two circles are very close, more so than some other contiguous circles like 2–3, 4–5, and 6–7. When the two circles are near to each other the position of their respective variables is similar to the one found for two behavioral circles. Any given variable correlates highest with the corresponding one in the other circle, so that the interrelationship between the two sets of variables should preserve the circumplex pattern. This pattern is indeed apparent in the matrix of Table 8, giving the intercorrelations between the variables of perceptual Circles 1 and 2 for the husband. In this matrix there are 11 deviations from the prediction. The matrix for Circles 7 and 8, also very near to each other, is similar to the matrix of Table 8 (see Appendix).

Let us now consider two circles which are quite apart in the figure like Circles 1 and 6, social acceptance of other and emotional rejection of self. Figure 3 suggests, for example, that Variable III of Circle 6 is nearer, or more related, to Variable VIII of Circle 1 than to its corresponding Variable III of the same Circle 1. In other words, the ideal self of the wife is more related to what she actually receives from her husband than to her norm of behavior toward her husband. The matrix of intercorrelations between the variables of Circles 1 and 6 (for the husband observer) is given in Table 9. Inspection of this matrix shows indeed that the above prediction is supported: the correlation between III, 6 and VIII, 1 is higher, .21, than the correlation between III, 6 and III, 1, .05. In general, in this matrix the highest correlations are around the border, and the correlations tend to decrease as one moves toward the center of the matrix. This intercorrelation pattern is similar to the one of Tables 4 and 5 for the multiple coefficients of correlation and is obviously different from the circumplex pattern. In a circumplex the highest correlations are

TABLE 8

AN EXAMPLE OF INTERCORRELATIONS BETWEEN TWO PERCEPTUAL CIRCLES NEAR TO EACH OTHER

	Perceptual Type	Behavioral Type 1							
		Perceptual Type							
		I	II	III	IV	V	VI	VII	VIII
Behavioral Type 2	I	72	57	51	36	29	44	55	57
	II	61	72	55	47	41	41	43	55
	III	57	59	70	52	48	45	46	50
	IV	48	54	56	61	39	42	37	44
	V	26	34	33	29	61	38	32	41
	VI	39	40	40	38	40	60	50	52
	VII	52	43	43	37	38	53	64	64
	VIII	56	54	44	36	39	47	57	76

TABLE 9

AN EXAMPLE OF INTERCORRELATIONS BETWEEN TWO DISTANT PERCEPTUAL CIRCLES

	Perceptual Type	Behavioral Type 1							
		I	II	III	IV	V	VI	VII	VIII
Behavioral Type 6	I	15	15	09	11	06	06	13	18
	II	14	13	12	11	04	02	14	14
	III	14	13	05	08	01	01	06	21
	IV	09	09	05	06	−01	05	06	12
	V	18	16	10	12	00	05	10	13
	VI	25	20	16	16	04	06	10	17
	VII	22	16	09	08	04	07	09	17
	VIII	25	18	15	15	04	09	12	17

found near the main diagonal; here they are found around the borders of the table and the correlations decrease as one moves toward the center of the table. The more interpersonal the two correlated variables, the higher their coefficient of correlation. The correlation pattern of Table 9 indicates that ideal-self-rejection correlates highest with lack of actual acceptance from the other. Actual behavior of the other toward the self influences both the actual and the ideal self, while the norm of the other appears to play a minor role. Actual behavior is more interpersonal than ideal behavior; the above results support, therefore, the prediction that the relationship between actual and ideal behavior should be higher than between different kinds of ideal behavior.

We have considered two extreme examples of relationship between perceptual circles. When the two circles are very near to each other, as 1 and 2, their intercorrelations approximate the circumplex pattern, each variable tends to correlate highest with the one that corresponds to it on the other circle. When the two circles are distant, as 1 and 6, the correlations pattern tends toward the single-factor model of Spearman (1927); in Spearman, the single factor is general intelligence, here it is the interpersonality of the two variables. A less interpersonal variable in one circle will correlate more with a strongly interpersonal one in the other circle than with its corresponding variable. So it happens, in the example given above, that ideal self is related to actual other more than to ideal other.

The ringex model suggests that the effect of interpersonality becomes stronger with the increase of the distance between the two intercorrelated perceptual circles. Therefore, it may be expected that, as distance increases, the correlation matrix will move away from the circumplex pattern toward the single-factor pattern.

This proposed interplay of interpersonality with the circumplex pattern may be illustrated by an example which is of some interest in terms of social exchange. Let us consider the correlation of Variable VII from one circle with Variables I and VII from another one. Variable I is more interpersonal than VII; if the correlation is determined by the interpersonality of the two variables then: $r(\text{I–VII}) > r(\text{VII–VII})$. The circumplex model, on the other hand, leads to the opposite prediction: Variable VII should be closer to the corresponding VII, on the other circle, than to I. Taking now the variables from Circles 1 and 4, which are rather distant, support the first prediction, interpersonality gains the upper hand. The second prediction, following the circumplex model, is however supported when the variables are taken from the two contiguous circles, 2 and 3. The respective coefficients, for the observer husband, are: $r(\text{VII, 4–I, 1}) = .38$; $r(\text{VII, 4–VII, 1}) = .30$; $r(\text{VII, 3–I, 2}) = 30$; $r(\text{VII, 3–VII, 2}) = .45$. For the wife the coefficients are closely similar.

When these results are translated into plain English, with the help of Table 1, they read as follows:

1. Giving oneself status (VII, 4) is more

related to receiving status from the other (I, 1) than to giving it to him (VII, 1).

2. Giving oneself love (VII, 3) is more related to giving love to the other (VII, 2) than to receiving it from him (I, 2).

These results suggest that the economics of status and love exchange are somewhat different. In love, the more one gives to the other the more he has for himself. To have status, however, one has to receive it from the other.

Several examples of intercorrelation matrices between perceptual circles are given in the appendix (see Footnote 3). In general they tend to follow the proposed pattern. There are, however, several deviations and also some systematic features which require further study. In analyzing matrices relating the behavior toward the other to the behavior toward self, as for example, the matrix of Table 9, one should remember that when the behavior is toward the other there is a gap between Types IV and V. When the behavior is toward the self the gap is between Types I and VIII. This has the effect of producing a certain shift in the relationship between the variables of the two circles.

The evidence presented in the earlier sections clearly supports the ringex model. The intercorrelations among perceptual circles are less conclusive to this regard and further analysis may lead to some modifications in the proposed model.

Relationship between Observers

So far we have been concerned with the relationship between variables as perceived by the same observer, the husband or the wife. It has been suggested that these variables are organized in a cognitive pattern which has been called the ringex. To gain some understanding of the relationship between the two observers, let us now intercorrelate the variables of a circle from the husband's ringex with the variables of the corresponding circle from the wife's ringex.

An example of the intercorrelations among the behavioral types of the two observers is given in Table 10. It refers to the actual behavior of the husband as perceived (by the two observers) from his point of view. This table shows how the husband's perception of his actual behavior toward his wife and toward himself is related to the wife's perception of the same behavior. Both observers take the point of view of the husband. The intercorrelation pattern of Table 10 tends toward the circumplex, in spite of the 10 deviations found in it. Some of the deviations are due to the fact that certain diagonal entries are lower than expected: this feature suggests a tendency toward the single-factor pattern. Indeed the relationship between the wife's perception of the husband's behavior toward himself and the corresponding perception of the husband is sometimes lower than expected in a circumplex, but still higher than in the single-factor pattern.

The effect of the interpersonality of the variables is more apparent in the intercorrelations among the perceptual types of the two observers, for a constant behavior type, given in Table 11. Type I for one observer corresponds to Type VIII for the other observer, Type II, to Type VII and so on. This correspondence can be checked by using the facet definition of these types in Table 1. Thus, to place the correlation between corresponding types in the main diagonal of the table the order of one observer (the wife) had to be reversed.

This table relates the perceptions of the two observers of their mutual social acceptance, actual and ideal. The coefficients tend to be higher at the four corners of the table and to decrease toward the center, following the single-factor pattern. Thus, the highest correlations are between the perceptions of actual behaviors. The norm of one observer, on the other hand, tends to be associated with the other observer's perception of actual behavior, at least as much as with the other observer's norm.

These results suggest that the relationship between observers tends to follow the same pattern as within one observer, but the distance being larger, as shown by the smaller correlations, the effects of interpersonality tend to become stronger than within the same observer. Apparently, each observer tends to infer the other's perception of less interpersonal behaviors, such

TABLE 10

INTERCORRELATIONS AMONG THE BEHAVIORAL TYPES OF THE TWO OBSERVERS (ACTOR: HUSBAND; LEVEL: ACTUAL; ALIAS: HUSBAND)

	Behavioral Type	Observer wife							
		1	2	3	4	5	6	7	8
Observer husband	1	32	24	12	08	−05	03	22	22
	2	26	21	09	04	05	00	20	20
	3	21	16	32	23	04	06	00	02
	4	08	05	23	19	08	10	−03	−02
	5	04	04	10	14	18	06	10	06
	6	11	08	05	06	13	13	13	14
	7	26	21	01	−01	03	14	37	30
	8	33	29	07	−02	01	09	29	32

as ideal behavior and behavior toward oneself, from the more interpersonal ones. More interpersonal behavior appears to be more visible or more overt than the less interpersonal one.

All this suggests that there are two ways for building an image of the private picture of the other. One possibility is to use as a point of departure our own private picture: then covert behavior will be related to covert behavior more than to overt behavior and this will produce a circumplex pattern. Another possibility is to infer the private world of the other from his overt behavior: then the relationship between covert and overt will be higher than between covert and covert behaviors; this will produce a single factor pattern. When both ways are used the pattern may be somewhere between these two ideal types. Some of our results also suggest that the observer's own self is influenced by the overt behavior of the other toward him, so that in this case the intercorrelations will again approach the single-factor pattern. The ringex model attempts to predict which pattern can be expected in each particular case.

Reliability of the Data

To assess the reliability of some of the present findings, one can compare them with the results of some earlier investigations. Some indirect support for the hypothesis of order of the perceptual types is provided by a study of the foreman-worker role (Foa, 1958), using the same perceptual facets as in this study, but an entirely different technique of observation: pictures rather than a questionnaire.

With regard to the circular structure of the behavioral types, there is quite a number of studies (for a review, see Adams, 1964, and Foa, 1961) pointing in the same direction. The above studies employed

TABLE 11

INTERCORRELATIONS AMONG THE PERCEPTUAL TYPES OF THE TWO OBSERVERS (BEHAVIORAL TYPE 1: SOCIAL ACCEPTANCE OF OTHER)

	Perceptual Type	Observer wife							
		VIII	VII	VI	V	IV	III	II	I
Observer husband	I	32	26	19	14	18	27	32	31
	II	31	28	25	17	23	24	30	26
	III	26	25	25	15	18	22	23	23
	IV	23	19	24	19	21	19	21	18
	V	17	14	18	12	22	20	22	19
	VI	20	20	19	17	20	23	26	26
	VII	22	16	12	10	19	28	32	33
	VIII	29	21	17	12	23	32	34	32

techniques of observation different from those used in this study. This will answer the possible criticism that the regularity of the results may be due to some artifact of procedure in gathering the data. There is nothing in the procedure used which adds factual support to such a suspicion. The only instance in which the order of the questions is identical with the order of the variables is in perceptual Types II–III and VI–VII, and these types do not correlate higher than other contiguous types which were observed by nonconsecutive questions.

With regard to the ringex structure as a whole, there is certainly need for further validation. The results obtained seem, however, to be good enough for accepting it as a point of departure for future investigations.

Discussion and Conclusion

We have attempted to present a picture of the cognitive organization a person has of his relationship to another person in reciprocal roles. A beginning has also been made in relating the picture of one person to the picture of the other one. The proposed cognitive organization is essentially based on two rationales: the developmental and the interactive ones. The ringex suggests which of the two will be prevalent in determining a particular interrelationship pattern. The developmental rationale proposes that the relationship among variables is determined by the manner in which these variables become differentiated during the psychosocial development of the child. We have seen that this rationale leads to the prediction of the order of contiguity of the variables: usually a circumplex order and, more rarely, a simplex one. The interactive rationale, on the other hand, proposes that the relationship between variables is determined by the interpersonal situation, here and now. According to this rationale the variables can be ordered from the most interpersonal or overt to the least interpersonal or covert. It follows that a covert variable will be more closely associated with an overt one than with another covert variable. In this context the notion of maturing may be understood as moving away from the organization pattern resulting from the development sequence, toward a pattern more attuned to the realities of a specific interpersonal situation.

This same issue of development versus interaction is also present in the organization of family roles (Foa, Triandis, & Katz, 1966). It has been suggested, for example, that the role of husband toward wife is modeled on the role of son to mother. To what extent will the husband's role remain similar to the son's role or will be influenced by the behavior of the wife toward the husband? The structure of the family roles suggests the behavioral influence may be stronger when the two reciprocal roles occupy similar power positions.

This problem is intimately related to the question of whether interpersonal behavior can be changed and how. Conditioning therapy appears to be close to the interactive viewpoint, as it is in any other "here and now" therapy. Psychoanalysis puts more stress on the developmental aspect: the developmental process has to be somehow reexperienced in order to obtain a change in behavior. Our data seem to indicate that both development and interaction influence behavior and may provide some pointer in suggesting which combination of techniques may be best for obtaining a certain behavioral change. It may become possible to do so when a number of research problems suggested by the present findings are solved. Some of these problems will be briefly outlined here below.

Further Research

Even a modest advance in the understanding of interpersonal behavior is unusually suggestive of new research. This may be due to the fact that interpersonal behavior is at the crossroad of several areas of psychology, relating, as it does, to child development and cross-cultural research, as well as to role analysis and clinical psychology. The role that interpersonal behavior may play in the theoretical

integration of these different areas can be exemplified by the following examples of research problems generated by the present findings.

1. The data of this study refer to a sample of a "normal" population of married couples. It becomes of interest to investigate how various kinds of psychiatric cases differ from normals and the effects of therapy on these differences. It has been suggested that the average frequency of types of behavior may differ (Adams, 1964). It is now proposed that such differences may also be found in the *size* of the correlation coefficients among variables, while the *order* position of the variables in the structure may remain invariant. The correlation between certain variables may be higher in data obtained from psychiatric cases than for normals in a given kind of disorder and lower than for normals in another kind of disorder. If some correlations increase (or decrease) as compared to normals, some other correlations will decrease (or increase) in the same kind of disturbance. It is proposed, in other words, that in psychiatric patients variables will maintain the same order as in normal individuals but will change their respective distance or degree of differentiation: some variables will approach certain others while moving away from other variables. Some findings of other investigators relating, for example, to differences between actual-self- and ideal-self-perception among normals and various kinds of mental patients seem to point in this direction. This type of investigation may ultimately lead to a new typology of behavior disturbances, based on structural differences, and relatively simple tests for differential diagnosis.

2. Changes in degree of differentiation within the same order may also be found in different cultures. Results supporting this hypothesis in two cultural groups are reported in another paper (Foa, 1964). These changes in differentiation from one culture to another one seem to be systematic and related to the value system of the culture. Some preliminary evidence suggests that cross-cultural differences in the differentiation of interpersonal behavior leads to tension and conflict when persons from different cultures have to cooperate in a common task or to negotiate. It has been found possible to reduce heterocultural strain by training a person from one culture to make the interpersonal differentiations that are required in the other culture (Foa & Chemers, 1966).

3. The ringex structure refers to a given pair of reciprocal roles. It appears of interest to investigate the relationship among the various roles of a person, both from a developmental and a structural point of view. It is commonly accepted that the different roles one has to play in adult life originate from a small number of roles developed in early childhood. It has often been suggested that one of the ways in which a child acquires new roles is by taking the role of the other. It is a common experience to see young girls playing the role of the mother toward a doll, while the doll is, so to speak, playing the role of the girl. This exchange of roles, which has been so far described intuitively, finds a precise expression in the ringex structure: the new role is created by interchanging the elements of the actor facet in the old role. The actor "nonobserver" of the old role becomes the actor "observer" of the new role and vice versa. Thus the new role will, at least at the beginning, look like the mirror image of the old one. These considerations have a direct bearing on the problem of ordering different roles in a contiguity pattern. A beginning in this direction has been made by a study of cross-cultural invariance in the organization of the roles of the family system (Foa, Triandis, & Katz, 1966). It remains now to investigate the organization of roles in other social systems such as the school, work, religion, and their relationship to family roles. The work done on the family roles indicates that the semiringex of one actor will be near to the semiringex of the other actor, as in our case, only when the two actors occupy similar power positions, as husband and wife. When the power is different, as in father and son, each one of these two roles relates more to some other role than to its reciprocal role. This suggests that, when dealing with all the roles of a social system, it may be more con-

venient to consider, as a unit of analysis, the semiringex of each role rather than the ringex of two reciprocal roles.

4. The sequence of differentiation of interpersonal concepts, which has been proposed here, needs to be tested through the investigation of the cognitive organization of children of different ages. Such a testing requires, however, some further theoretical work. Two sequences have been presented here: one for the behavior facets and the other for the perceptual facets. A third sequence, differentiating between the facets of family roles (and having the actor facet in common with the perceptual sequence given here) has been described by Foa, Triandis, and Katz (1966). Obviously these three sequences, involving eight facets, have to be brought together in a single development pattern which may account for the organization within and between family roles.

Some findings (Foa, 1964) suggest that differentiation at a given stage of the sequence may be more or less strong depending on the value system of the culture of the child. The culture is, of course, mediated to the child through his immediate social environment, the family, and peer groups. It becomes then of importance to know how efficient is the environment in producing in the child the degree of differentiation prescribed by the culture. If too much or too little differentiation occurs in the cognitive development of the child this may result in maladjustive interpersonal behavior. The possibility that behavioral problems may be related to deviations from the degree of differentiation, which is normal in a particular culture, has been discussed earlier.

The theoretical approach underlying these proposals may be summarized as follows:

a. The sequence of differentiation in the child and the resulting cognitive organization of interpersonal behavior in various roles of the adult are cross-culturally invariant.

b. The degree of differentiation at a given stage is prescribed by the culture and may thus vary from one culture to another one.

c. When, as a result of social environmental conditions in childhood or other factors, the differentiation actually made is lower or higher than the culturally prescribed one, and thus differs from the one of most other individuals, maladjustive behavior may occur.

d. A similar maladjustive situation occurs when an individual is operating in a culture other than his own. Here, however, his cognitive organization differs from the prevailing one because his culture is different and not because he deviates from his own culture.

These considerations suggest that the study of the differentiation and organization of interpersonal behavior in different roles may provide a focus for the theoretical integration of concepts from such different areas as developmental, cross-cultural, and abnormal psychology.

REFERENCES

Adams, H. B. Mental illness or interpersonal behavior? *American Psychologist,* 1964, **19,** 191–197.

Foa, U. G. The contiguity principle in the structure of interpersonal relations. *Human Relations,* 1958, **11,** 229–238.

Foa, U. G. Convergences in the analysis of the structure of interpersonal behavior. *Psychological Review,* 1961, **68,** 341–353.

Foa, U. G. The structure of interpersonal behavior in the dyad. In J. Criswell, H. Solomon, and P. Suppes (Eds.), *Mathematical methods in small group processes,* Stanford: Stanford University Press, 1962. Pp. 166–179.

Foa, U. G. A facet approach to the prediction of communalities. *Behavioral Science,* 1963, **8,** 220–226.

Foa, U. G. Crosscultural similarity and difference in interpersonal behavior. *Journal of Abnormal and Social Psychology,* 1964, **68,** 517–522.

Foa, U. G. New developments in facet design and analysis. *Psychological Review,* 1965, **72,** 262–274.

Foa, U. G., & Chemers, M. M. The significance of role behavior differentiation for cross-cultural interaction training. Technical Report No. 22, Contract DA-49-193-MD2060. Urbana, Ill.: Department of Psychology, University of Illinois. 1966.

Foa, U. G., Triandis, H. C., & Katz, E. W. Cross-

cultural invariance in the differentiation and organization of family roles. *Journal of Personality and Social Psychology*, 1966, **4**, 316–327.
GUTTMAN, L. A new approach to factor analysis: The radex. In P. R. Lazarsfeld (Ed.), *Mathematical thinking in the social sciences.* Glencoe, Ill.: Free Press. 1954, Pp. 258–348.
GUTTMAN, L. What lies ahead for factor analysis? *Educational and Psychological Measurement*, 1958, **18**, 497–515.
PIAGET, J. *The construction of reality in the child.* New York: Basic Books, 1955.
SPEARMAN, C. *The abilities of man.* London and New York: MacMillan, 1927.

(Received September 17, 1965)

Vol. 80, No. 16 Whole No. 624, 1966

Psychological Monographs: General and Applied

INTELLECTUAL ABILITIES OF SYMBOLIC AND SEMANTIC JUDGMENT[1]

RALPH HOEPFNER, KAZUO NIHIRA, AND J. P. GUILFORD

University of Southern California

2 studies approached the problem of describing judgmental processes from the standpoint of individual differences in terms of basic traits. Based upon Guilford's structure-of-intellect model, the factors of symbolic and semantic evaluation were hypothesized to exist as distinct from one another and also from factors represented in other domains of the model. Experimental tests were developed as measures of the hypothesized factors. Measures of reference factors were also employed to demonstrate the uniqueness of the hypothesized factors. The tests were administered to 2 samples of high-school students, scores were factor analyzed, and axes analytically rotated, resulting in the demonstration of the 12 hypothesized evaluation factors and all the reference factors as uncorrelated dimensions of intellectual ability. The conclusion is that the model has continued to lead fruitfully to undiscovered, differentiable intellectual aptitudes.

FOLLOWING several successful attempts to validate Guilford's theory of intelligence, in which intellectual aptitude factors of creative production, problem solving, and symbolic thinking were isolated and investigated, the attention of the Aptitudes Research Project turned to the intellectual operation of evaluation. Other operations—operations being one of the three facets of Guilford's model—had received attention previously; divergent production in several studies of creative potential; and cognition and convergent production in studies of problem-solving and symbolic factors.

In the most recent explication of his complete model, Guilford and Merrifield (1960) define the operation of evaluation as reaching decisions or judging on the basis of goodness according to certain criteria. Although inclusion of this operation in the model of intelligence was somewhat theoretical, there was a history of discovery of evaluation-like factors, defined by tests that did not correlate highly with cognition and production tests, but which formed weak factors of their own, instead.

A History of Evaluation Factors

The first of these factors can be attributed to L. L. Thurstone (1938a) when he identified the factor that became known as "perceptual speed," and that later was recognized as an evaluation factor—the evaluation of figural units (EFU). The factor has been most consistently defined by tests that require the rapid comparison of figural objects with judgments of identity versus nonidentity (Guilford & Lacey, 1947). But there have been times when there was uncertainty as to whether it also applied to tests requiring the identification and matching of letters, numbers, and words (Coombs, 1941; Thurstone, 1938b). Thus, the perceptual-speed factor is of interest, since the structure-of-intellect model forecasts a distinct but parallel ability concerned with the identity versus nonidentity of literal material, and tests of EFU could serve as models for the hypothesized evaluation of symbolic units (ESU) factor, which heretofore had not been

[1] The studies reported herein are two in a series conducted by the Aptitudes Research Project at the University of Southern California, under Contract Nonr-228(20) with the Office of Naval Research, Personnel and Training Branch. The ideas expressed do not necessarily reflect the views of that agency. The authors wish to extend their special thanks to Philip R. Merrifield, who aided greatly in the planning and the supervision of the construction and administration of the test batteries.

clearly demonstrated. One or two studies cited by French (1951) gave some hope of such differentiation, for example, Bechtoldt (1947), but there was nothing decisive.

The factor of "judgment" was found in the Army Air Force research during World War II (Guilford & Lacey, 1947). It was largely identified by its association with a test called Practical Judgment, which was composed of verbally stated problems or predicaments of a common everyday type. Multiple-choice answers offered different more-or-less plausible solutions, the examinee (*E*) to select the best alternative. A similar factor has been identified as judgment in each of three studies by the Aptitudes Research Project (Berger, Guilford, & Christensen, 1957; Guilford, Green, Christensen, Hertzka, & Kettner, 1954; Kettner, Guilford, & Christensen, 1959a), sometimes with Practical Judgment as a marker test, sometimes not.

"Speed of judgment" was first identified as a factor by Thurstone (1944) in a study of perception. The tests which defined this factor involved time scores of making choices of color versus form (in classifying figures), of desirability of personality traits, and of weights involving the size-weight illusion. The perceptual nature of this factor indicated its resemblance to a factor called "perceptual evaluation," isolated later in a factor-analytic study of evaluative abilities by Hertzka, Guilford, Christensen, and Berger (1954). The leading test for this factor involved the ability to judge rapidly the length of pairs of lines or the size of given figures. Thus, the factor was defined as the ability to appraise rapidly the similarities and differences among simple perceptual materials. Both "speed of judgment" and "perceptual evaluation" factors are primarily concerned with this ability. It seems that these factors are quite similar to the well-known factor of "perceptual speed," which has to do with judgments of identity versus nonidentity of figural material. The authors are not aware of any systematic studies which investigated the equivalence or differences among those factors of perceptual evaluation.

In the same factor-analytic study (Hertzka et al., 1954), a factor called "speed of evaluation" was isolated. The leading tests for the factor involve judging whether or not named objects satisfy simultaneously certain specified criteria, such as roundness and hardness. In such tests, decisions are easy, so that speed becomes an important variable, hence the naming of the factor. Prior to Hertzka's study, Bechtoldt (1947) independently identified a similar factor which was later called "speed of association" by French (1951). The tasks involved in the tests that defined the speed-of-association factor are essentially conceptual or verbal rather than perceptual evaluation, since the tests deal with the meanings of words, objects, and sentences. Thus, the speed-of-association factor is probably equivalent to Hertzka's speed-of-evaluation factor.

The factor of "logical evaluation," also formerly called "logical reasoning," has appeared in a number of analyses (Frick, Guilford, Christensen, & Merrifield, 1959; Green, Guilford, Christensen, & Comrey, 1953; Guilford et al., 1954; Hertzka et al., 1954; Kettner et al., 1956). The most consistent tests identifying the factor were in the form of multiple-choice syllogisms and variations of tests in the syllogistic category. The factor has been considered evaluative because in these tests *E* is not required to produce conclusions but to evaluate conclusions presented to him. There was a seemingly parallel factor, identified most consistently by a test called Symbol Manipulation, which was essentially syllogistic in form but with letter symbols rather than words. This is the nearest that research has come to demonstrating previously a symbolic-evaluation factor.

The factor of "experiential evaluation" has been defined chiefly by reference to the test that has most strongly represented it in two analyses—Unusual Details (Hertzka et al., 1954; Marks, Guilford, & Merrifield, 1959). The test presents pictured situations, with *E* to state two things he sees wrong with each picture. The "things wrong" may be contrary to past experience or to other items of in-

formation within the pictures. The interpretation and naming of the factor have placed emphasis upon the obvious role of past experience, without much consideration that most judgments (and other mental operations as well) depend upon past experience.

"Sensitivity to problems" was first hypothesized as an important ability contributing to creative thinking (Wilson, Guilford, Christensen, & Lewis, 1954) and was demonstrated as a factor. It has since been verified as a factor in other studies (Kettner et al., 1959a, 1959b; Marks et al., 1959; Merrifield, Guilford, Christensen, & Frick, 1962). The factor was defined as the "ability to see defects, needs, and deficiencies," and was thought to belong in the evaluative domain, with the conjecture that simply being aware of the existence of problems, or aware that things are not all right, is an instance of evaluation (Guilford, 1959).

DEFINITIONS

The review of these known factors enabled the writers to draw some inferences that were logically helpful. The two major deductions have to do with the fact that evaluation is a multiple-dimensional affair, and with the nature and definition of evaluation. With regard to the multidimensional hypothesis, the indication is that there are as many as five verbal- or semantic-evaluation factors. The same hypothesis is supported by the recognition of as many as three nonverbal evaluation factors. Thus, the way seemed open for a considerable number of differentiable evaluative abilities.

The conclusion concerning the definition of evaluation itself could not be so univocal as that concerning the existence of multiple factors. An empirically based definition is best achieved from examination of the factors and of their tests. As the factors discussed above are considered, one finds that the crucial type of activity can be variously described. Perhaps the most general property of evaluation tests is that they require decisions. Multiple-choice tests require decisions among alternative answers. Other tests require essentially yes-no decisions. As many testers well know, a multiple-choice test can also involve a set of yes-no decisions, with each alternative answer (together with the item stem) being a true-false item. In a multiple-choice syllogism test, either of verbal or symbolic content, *E* may make a yes-no decision regarding every alternative answer; it does or it does not follow logically from the premises.

Still other factor tests emphasize sensitivity, in which *E* has to detect defects, or deficiencies. The tests usually are in completion form, and no choice is offered among alternatives. The sensitivity interpretation can be applied meaningfully in syllogistic tests, in which detection of logical errors or discrepancies is needed. It might be applied to tests involving yes-no decisions regarding identity of pairs of objects or series of symbols—a sensitivity to differences or discrepancies—and to tests involving the criterion of satisfaction of specified criteria, in the form of detection of failure to meet the criteria.

The issues just raised were not fully realized or fully met at the time part of this study was planned, but they were given attention in connection with the study of symbolic-evaluation abilities. It turned out that in the semantic study there was a bias in favor of the hypothesis that evaluation involves decisions among alternatives, and most tests selected or developed for that study are of multiple-choice form. Thus, relative judgments were emphasized rather than yes-no judgments or detection of things wrong. Several kinds of criteria were recognized and utilized, however, including identity, suitability, or effectiveness of given information for certain purposes.

In formulating a definition of "evaluation" for the symbolic study, several alternatives were considered. Two of them will be mentioned, since they have a direct bearing on the distinction between "sensitivity" tests and "estimation" tests. We shall then state a broader definition that embraces both conceptions.

In a definition that equates evaluation

to "sensitivity to error," the term "error" is interpreted broadly to include any kinds of defects, deficiencies, departures, inconsistencies, incongruities, etc. This view implies absolute judgments: to each individual a thing can be judged as being right or identical with another (within tolerance limits upon those variables thought to be relevant), or it is not. Some individuals detect such "errors" with low tolerance for deviations with respect to relevant information where others cannot do so. This is not to say that there is a dichotomy of individuals; they can still vary by small degrees along a continuum of greater or less sensitivity.

In the definition that emphasizes "estimation," it is implied that individuals also make relative judgments. When items of information fall short or deviate from a standard, one may deviate farther than another. It may be obvious that all the items of information depart from the standard, but which one deviates least? Where sensitivity tests typically call for absolute judgments of a yes-no, disjunctive type, estimation tests typically offer alternative items of information and ask which one deviates least, or sometimes (but rarely) which one deviates most. A ranking of alternatives is implied. This view of estimation is concordant with that offered by Johnson (1955).

Actually, the two views can be brought logically under the same definition of evaluation. In both cases, a standard of some kind is implied. In both cases, criteria for judgment are implied. A definition embracing both views would read: Evaluation is a matter of decision concerning criterion satisfaction. This is the definition adopted as the basis for planning these studies. One of the objectives was to determine whether tests embodying the sensitivity principle and those embodying the estimation principle would both indicate the same kinds of ability, thus justifying subsuming them both under a single definition of evaluation.

When it is said that evaluation is concerned with criterion satisfaction, it is then necessary to give attention to what kinds of criteria are suitable for use in tests of evaluation abilities. Some of the traditional criteria have been: identify versus deviation, completeness versus incompleteness, compatibility versus incompatibility, congruity versus incongruity, effectiveness versus ineffectiveness, and suitability versus unsuitability. For testing purposes, a criterion must be of a type that can be communicated to the examinee. As will be seen in discussions of the results, some additional kinds of criteria were included in these studies, such as popularity versus unpopularity (frequency of usage) and highly probable versus improbable. With symbolic and semantic materials to be evaluated, the use of either esthetic or moral criteria can be and was avoided. Questions regarding those two kinds of criteria are more likely to arise with greater urgency in analyses pertaining to figural information on the one hand and behavioral information on the other.

Inferred in the preceding discussion of evaluation is the fact that the study of evaluation abilities was carried out in two parts, a symbolic study (Part 1) and a semantic study (Part 2). Such a split program was necessary due to the amount of testing time available for any one group of examinees.

The Expected Factors and Their Tests

During the late 1950's there was developed a theoretical model of intellectual abilities called the "structure of intellect" (Guilford & Merrifield, 1960), designed to bring into a single systematic classification the intellectual factors then known. In the model the primary intellectual abilities are classified in terms of three major dimensions: operations, contents, and products. There are five kinds of operations which the organism can perform: cognition, memory, divergent production, convergent production, and evaluation. There are four kinds of contents, or broad categories of information, upon which the organism can perform the operations: figural, symbolic, semantic, and behavioral. The third dimension of the model represents the six kinds of products: units, classes, relations, systems, transformations, and implications. The products are the re-

sults of the organism's psychological processing of information.

The six kinds of products of the model have been defined in a number of places, but will be very succinctly defined here. Units are segregated items of information having thing character. Classes are groups of items of information having common properties. Relations are meaningful connections between units. Systems are organized or structured complexes of units and relations. Transformations are changes or redefinitions of known items of information. Implications are in the form of expectancies, predictions, or consequences of information.

An objective of some importance in these studies is that they represent further attempts to validate Guilford's model as a source of hypotheses for defining and isolating factors of human intelligence. If the model continues to generate concepts which can be found to represent unique abilities, then its contribution to psychological explanation and prediction will be further substantiated.

Not only is the existence of new factors deduced from the model, but the model further offers operational specifications for the measures needed for the factors. Such measures, which serve as the "empirical world" for verification of the factors and the model, are then available as instruments for study of the traits in new investigations involving those traits. Today's new factors and their experimental tests become tomorrow's reference concepts and marker tests for use in other kinds of investigations. They also become available for applied psychological prediction and selection, which is the ultimate social-value testing of the model itself.

The studies reported here attempted to explore intensively the 12 factors of symbolic and semantic evaluation hypothesized from the model. The hypothesized factors were deduced from the model, and the hypotheses were to be tested by factor analysis. Previous studies at the Project have shown the fruitfulness of such a research strategy.

Since most test responding can be viewed as a problem-solving process, it is worthwhile to consider the nature of evaluative processes in terms of a problem-solving model. Such considerations will lead to a rationale for developing the operational specifications for separating evaluative abilities from the rest of the problem-solving process.

According to Dewey's (1910) problem-solving model, evaluation plays its most important role at the last phase of the problem-solving process. In this sense evaluation is used to mean the testing of the possible solutions produced by the problem solver before employing one of them.

In the structure-of-intellect model, evaluation is defined broadly as the process of decision as to whether any item of information that is cognized, remembered, or produced meets a certain standard or goal. It terms of the traditional problem-solving models, the operation of cognition corresponds to the phase commonly referred to as "understanding the problem." The operation of production, either divergent or convergent, corresponds to the phase commonly described as "suggestion of possible solutions." The operation of memory is required at all five stages of Dewey's model, since stored information has some bearing upon every mental event.

The operation of evaluation should play an important role at the last phase of the problem-solving process if the possible solutions are at all doubtful or competitive. But evaluation may also be called for at any phase of the process, whenever there is uncertainty as to whether the information, either remembered, cognized, or produced, meets specified criteria.

The problem of the operational specification for separating evaluative abilities from parallel cognitive abilities deserves a few comments. In the preceding section, evaluation was defined as the process of deciding whether information that is cognized, remembered, or produced meets certain standards or goals. A test designed to measure an evaluative ability should pitch the evaluative aspect of the task at a level of difficulty that will ensure that individual differences in test scores reflect that source. On the other hand, whatever cognitive, memory, or

production aspects the test may have should be made so easy that they contribute nothing to variance of the scores. Difficulty can be introduced into the decision process by providing an appreciable degree of uncertainty as to which of several alternatives is best. This can be achieved by making all of the alternative answers relevant, even acceptable under low standards of acceptability, and about equally desirable. Cognition can be kept at a low level of difficulty by specifying the problem and the criteria very clearly and by using very familiar information.

In designing factor-analytic studies, a sufficient number of tests for each hypothesized factor should be present in the test battery, so that each factor axis may be overdetermined in rotations and each factor clearly interpretable in terms of the apparent unique psychological function shared by all its tests and by no others. In this study, at least three tests were employed for each of the 12 hypothesized evaluation factors.

An important goal for the well-designed factor-analytic test of factor hypotheses is not only to determine what the experimental constructs are but also what they are not. The formal hypothesis states that the 12 factors of evaluation are not only distinct from one another, but are also distinct from other factors deduced from the model. For this reason, a number of marker tests, known from previous experience to measure reference factors, were included in the analyses to demonstrate the distinctness of the new experimental factors from factors already known.

The usual test of the distinctness of experimental factors is made by selecting for simultaneous analysis reference factors that might possibly be identical with the experimental factors. Of all the nonevaluation factors, those of cognition were suspected of being most likely to be confused with the experimental evaluation factors. One reason is that it takes care to construct an evaluation test that does not offer necessary cognition problems of sufficient difficulty to introduce some cognition variance into the total scores, or, indeed, that does not become a cognition test instead of an evaluation test. For this reason, tests of parallel cognition factors were selected as the most important marker tests in the analyses. In addition, tests of memory, divergent production, and convergent production were included when it appeared important to do so.

The Symbolic Study

The whole logic of symbolic communication as compared with conceptual or semantic communication is that more precision can be had due to the denotative inflexibility of symbols. One might then ask the question, "What is there to evaluate in connection with symbolic information?" We might expect some different aspects to evaluation of symbolic information than those that apply to semantic information, which is relatively rich with connotative meaning.

Several different varieties of information conform to the definition of symbols stated by Guilford and Hoepfner (1963): "Information in the form of signs, having no significance in and of themselves. [p. 2]." The clearest example of a symbol is a number. Numbers have no significance in and of themselves, yet can be evaluated for numerical identity, order, or consistency, with respect to other numbers. Letters also conform to the definition when they are processed in terms of their literal properties rather than their figural properties. Syllables can be symbolic units, as well as words, when their semantic meanings are not relevant to the task, as in breaking words into syllables or in word compounding. All these types of symbols were used in various experimental tests in this study.

The existence of six distinct product factors of symbolic evaluation provides the major problem of this study. The demonstration of these six factors, or the failure to do so, would confirm or fail to confirm the model from which the hypothesis was deduced. The tests designed as measures for each of the products will be described in detail later.

With three kinds of symbols available and with the distinction between sensitiv-

ity and estimation tests, for a completely systematic experimental design it would have been desirable to have six experimental tests for every product factor. No effort was made to achieve fully this kind of coverage with experimental tests, and it was difficult to achieve all six kinds of tests with every product. There proved to be enough dispersion of the conditions to make possible answers as to whether both sensitivity tests and estimation tests serve to measure these evaluation factors and whether the kind of symbol makes a difference in the success of tests.

Symbolic reference factors. The marker tests for reference factors selected for analysis in the symbolic study are described below. More complete descriptions for all the tests employed in this analysis can be found in Hoepfner, Guilford, & Merrifield (1964).

CSU—Cognition of symbolic units (symbol cognition):

Disemvowelled Words—Recognize familiar words with dashes in place of vowels; then complete the words by writing the vowels.

Word Combinations—Produce a new word from the ending of one word and the beginning of another.

CSC—Cognition of symbolic classes:

Number Classification—Select one of five alternative numbers to fit into each of four classes of three given numbers each.

Number-Group Naming—State how the numbers in each set of three are alike.

CSR—Cognition of symbolic relations:

Seeing Trends II—Describe a trend based upon relations of letters in a group of words.

Word Relations—Recognize the same relation between words in each of two pairs, then complete a third pair from five alternative words using the same relation.

CSS—Cognition of symbolic systems:

Circle Reasoning—Discover the principle by which one circle is blackened in each of four rows of circles and dashes. Apply the rule to the fifth row.

Letter Triangle—Find the pattern of the letters arranged systematically within a triangle.

CSI—Cognition of symbolic implications:

Symbol Grouping—Rearrange scrambled symbols in a specified systematic order as efficiently as possible.

Word Patterns—Arrange a list of short words efficiently in a crossword-puzzle design.

CMU—Cognition of semantic units (verbal comprehension).

Iowa Tests of Educational Development—Test 8, General Vocabulary—Recognize the meanings of words commonly used in communication. This test is similar to standard verbal comprehension (CMU) marker tests.

Preliminary Scholastic Aptitude Test—Verbal—PSAT is an abbreviated form of the SAT. The verbal score is the sum of scores on four tests: Opposites, Sentence Completion, Analogies, and Reading Comprehension. The dominant saturation is hypothesized to be CMU with some CMR variance contributed by the Analogies test.

Cooperative School and College Ability Test—Verbal—Verbal score is composed of scores on two tests, Sentence Understanding and Word Meanings. Tests similar to each have previously loaded on the verbal-comprehension (CMU) factor.

MSI—Memory for symbolic implications (numerical facility):

Numerical Operations—Rapidly add, subtract, or multiply simple numerical problems and select one of six alternatives as the answer.

DSC—Divergent production of symbolic classes:

Number Grouping—Group given numbers into several different classes based upon properties they have in common.

Varied Symbols—Find the different common properties that sets of letter combinations may have in common.

NSS—Convergent production of symbolic systems:

Operations Sequence—Produce the correct order of three specified numeri-

cal operations in order to get from one given number to another.

Word Changes—Arrange a list of words, each containing the same number of letters, so that the first word is changed into the last word with only one letter change at each step.

NST—Convergent production of symbolic transformations:

Camouflaged Words—Find within a meaningful sentence a group of consecutive letters that spell the name of a sport or game.

Word Transformation—Separate letters of words in a phrase with vertical lines to make a different set of words.

NSI—Convergent production of symbolic implications:

Form Reasoning—From the table, find the form that is implied by the three given forms.

Sign Changes—Solve simple arithmetic problems in which the operation sign is changed according to a set of rules.

EFU—Evaluation of figural units (perceptual speed):

Identical Forms—Find one of five figures that is exactly the same as the given figure.

Perceptual Speed—Rapidly match each of five objects to one of four given objects.

Finger Speed:

Marking Speed Test—Make as many Xs as possible in the rows of squares provided.

Symbolic evaluation factors. The tests for the hypothesized factors were developed using either of two approaches or a combination of the two. In the first approach, specific examples of tasks are deduced from the operation-content-product combination being investigated. For example, the ability in the cell EST, evaluation of symbolic transformations, involves evaluation of changes in symbolic materials. A code can be an example of a symbolic change, and so the test, Decoding, was developed.

The second approach emphasizes tasks similar to those for established factors having one or two attributes in common with the new factors. For example, ESU, evaluation of symbolic units, and EFU, evaluation of figural units, differ only with respect to the content category; test formats might be very similar.

Nearly 30 different pretest booklets were administered to classes in psychology at several colleges in the Los Angeles area. These pretestings were designed to obtain technical information such as the appropriate level of item difficulties, comprehension level of the test instructions, test reliabilities, and optimal time requirements for newly developed tests. Extensive item analyses were conducted whenever pretesting information revealed low reliability estimates.

From the reliability and intercorrelation data obtained from pretesting, 25 tests were selected from a pool of over 40 tests especially designed or adapted to measure the six experimental factors. The selected tests had pretest reliabilities in the .70's and .80's. Within-factor intercorrelations were generally considerably higher than between-factor intercorrelations, further ensuring the demonstration of the distinctness of the expected factors. The criteria of high reliability and desirable correlational pattern determined which tests were finally selected to represent the experimental factors in the final analysis. In the following paragraphs, the six experimental factors and the tests selected to measure them are discussed in detail.

ESU—Evaluation of symbolic units. The five experimental tests developed to measure evaluation of units had in common symbolic stimuli that are processed as wholes, rather than separated, analyzed, or classed. Although similar symbolic stimuli are employed in tests of the other intellectual products, the mental process performed upon units must maintain the thing quality of the stimuli. Guilford and Hoepfner (1963) had suggested tests employing letters and digits as stimuli for measures of ESU based on tenuous prior studies (French, 1951). Construction of the experimental ESU tests was based upon the history of parallel tests of symbolic units and the tryout of new kinds of stimuli.

The test, Correct Spelling, employed complete, common English words as sym-

bolic stimuli. The words function as symbols because *E* is to direct evaluation toward spelling rather than meaning. The *E* is tested on his sensitivity to the correctness or incorrectness of the spelled symbolic unit. In this case, sensitivity to spelling is based largely upon the long-term retention of the correct symbolic elements of standard English words. The words employed as items were selected from lists of commonly misspelled words published in English handbooks and secretarial manuals.

Derived from the format of a test used by Thurstone (1938a), Derivations also employs complete English words as test stimuli. Whereas Thurstone's test has *E*s make as many short words as they can in a limited time from the letters in a large given word, Derivations supplies not only the given word, but also 50 short words possibly derived from it. The *E*'s judgments are based upon sensitivity to the errors in some words that could not be derived from the long given words.

Familiar Letter Combinations is an experimental test that has a completely new type of symbolic stimuli: three-letter syllables. The *E* is to estimate which of two given syllables is more common as a part of real English words. Familiarity is the criterion for decision. Neither the syllables nor the criteria of real words are to be considered semantically; only the relative frequencies of occurrence are relevant. The key for this test was determined from the empirical frequency counts reported by Underwood and Schulz (1960). The nonsense syllables are paired so that the keyed syllable is far more commonly used than its alternative.

Letter "U" is a test of *E*'s sensitivity to the presence of a specified letter in words under speeded conditions. It is based upon Thurstone's test Letter "A" (1938b), which split its variance among factors that Thurstone called perceptual, number, and word factors. Bechtoldt (1947) found Letter "A" to be loaded highly on a factor with a test of crossing out specific letters on a page of regularly spaced letters. Although Cattell names the factor on which this test is loaded "speed of symbol discrimination" (Cattell, 1953), and Guilford and Hoepfner (1963) suggest the factor is ESU, French, Ekstrom, and Price (1963) conclude that our knowledge concerning this factor is not at all clear since several "subfactors" tend to pull together in different ways, depending upon the tests included in the factor-analytic battery. In general, a test like Letter "U" is often found in strong relation to perceptual-speed tests. To clarify this ambiguity, not only were four tests designed to measure ESU included in the battery along with Letter "U," but also two strong perceptual-speed (EFU) tests.

The test Symbol Identities was designed as a measure of *E*'s sensitivity to the identity or nonidentity of paired sets of numbers, letters, and words, under speeded conditions. It is essentially parallel to tests of EFU, in which identity of pairs of figures is in question. Symbol Identities is the only ESU test employed that cuts across all the possible stimuli considered appropriate in the symbolic domain.

Symbol Identities is similar to many of the tests designed to measure clerical speed and accuracy; *E* decides whether or not the two members of pairs of symbol sets are the same or different. This test, like Letter "U," could conceivably share much figural variance, as *E*s could compare each symbol stimulus, figure by figure, and arrive at an accurate judgment. Such activity is very inefficient, however; a figural attack upon Symbol Identities should result in poor performance, unless it is used only when a quick symbolic attack does not yield a decisive choice.

ESC—Evaluation of symbolic classes: Four experimental tests were developed to measure the factor ESC. A symbolic class was defined for this investigation as a group of symbols with some common property. Such a group of symbols would be composed of at least two members whose common property must be symbolic, not figural or easily semanticized.

In Best Number Class, *E*'s task is to assign given numbers to one of four classes in such a way as to maximize each number's value by assigning it to the most exclusive class it fits. The four classes into

which the stimuli were to be assigned were, in order of exclusiveness: EVEN MULTIPLES, ODD MULTIPLES, SQUARES, and PRIMES. The *E*s were warned that the numbers could possibly be assigned to several classes and that credit could only be earned by assigning each number to its most exclusive class.

The test Best Number Pairs is the other hypothesized ESC measure employing numbers as stimuli. The *E*'s task is to choose one of three pairs of numbers that makes the best class. In order from best to poorest, the classes are: pairs of perfect squares, pairs of multiples of the same number, pairs of odd or even numbers, and pairs with no class property.

Sound Grouping is a test with a long history. In each item, four words are given, three of which are fairly good rhymes and one is not. The latter is to be noted and selected, for the right answer. The test's factorial composition has been open to considerable question because of its tendency to go with different factors, depending upon the battery in which it has been analyzed.

Because of this history of factor instability and the fact that the previous studies did not include tests of what would now be called symbolic classes, Sound Grouping was hypothesized to measure ESC, a seemingly logical place for it. It was not expected, however, that Sound Grouping would suddenly become a unifactor test when placed in a battery with several ESC tests.

The fourth test designed to measure ESC is Word Choice. The *E* is to choose the best of three possible additions to a class of three words. The class properties used in Word Choice are symbolic, for example, order or nearness of certain letters or types of letters in the words. This test differed from other ESC tests in that none of the alternative words for any class completely possessed all the class properties; a best word had to be chosen, even though it was slightly wrong. It is thus an estimation test.

ESR—Evaluation of symbolic relations. The four tests developed to measure the factor ESR employed recognized connections, based upon symbolic variables, between symbolic units. Examples of connections based upon symbolic variables are "greater than," "equal number of consonants," and "similar ratios."

The first experimental ESR test, Related Words I, was adapted by analogy to Matched Verbal Relations, designed for factor EMR, evaluation of semantic relations. In Related Words I, *E* estimates which of three alternative word pairs is most similar to a given related pair. The relation between members of any pair is based upon the order and position of letters and the vowels and consonants that are changed or moved. This is the only ESR test in which no alternative answer is completely correct; only a best alternative is to be selected.

Sign Changes II had been developed as an ESR test to be used in a predictive battery for success in ninth-grade mathematics courses (Guilford, Hoepfner, & Petersen, 1965). The task in this test is to determine what sign changes, if any, must be made to change a numerical expression into an equation. An elementary understanding of arithmetical operations and the relationships of equality and inequality of expressions is all *E* needs in order to understand clearly the test items and the task.

Similar Pairs is a new test, in both idea and items. The stimuli are word pairs, the members of which are related by letter locations and letter changes. The *E*'s task is to judge whether the members in two such pairs are or are not similarly related. The process involved in responding to this test is sensitivity to sameness or differentness of the relations within the word pairs. In all the items, the relations within the pairs were kept extremely simple, so that there would be little or no difficulty in cognizing the relationships, so that cognition variance would be minimized in the test scores and the relative importance of the evaluation variance would be maximized.

Symbol Manipulation is a test of the ability to decide whether a given relationship between two letters follows logically

from other statements of relationship involving the same letters, where the relationships are "greater than," "equal to," and "less than," and their negations, all statements in symbolic form.

ESS—Evaluation of symbolic systems: Like the tests hypothesized to measure ESU, tests for ESS seemed to be easy to construct. Almost one dozen tests were developed to measure ESS and were pretested. Most of the tests at this experimental stage proved to have reasonably good reliability and reasonable intrafactor correlations. The five tests chosen to define the systems factor in the final analysis broadly cover the various types of symbolic content and sensitivity versus estimation.

All the stimuli for tests of symbolic systems are organized aggregates of units or relations wherein the interrelated or interacting parts are symbolically defined within the aggregate. The system, then, is the organization or pattern of parts, which may be compared with another system as to identity or similarity or which may be evaluated for internal consistency.

Best Letter Set was designed as a measure of *E*'s ability to estimate which of three sets of three or four letters each is most like a given set. The criterion of similarity is based upon the order and kinds of letters within the set. Although such small sets of letters might appear to function as units, the systems qualities of the alternative sets were sufficiently similar to force *E* to focus on them. It seemed highly unlikely that even the most sophisticated *E* could treat the stimuli as units and obtain a high score on this test.

Both Correct Letter Orders and Correct Number Series are tests of *E*'s sensitivity to internal inconsistencies in symbolic systems. The stimuli in both tests are sequences of symbols organized according to some simple principle, similar to items in familiar number-series tests. The systematic principle is stated verbally and *E* is to judge whether or not the sequence follows that principle.

The test Series Relations might also be considered an evaluative form of a number-series test, even though the task appears to be quite unlike that for Correct Number Series. In Series Relations, *E* is given a series of three numbers and is told that each element of the series except the first one is determined from the previous element (one to the left) according to some unknown rule. The *E* is then to estimate which of three alternative rules or operations would best relate each series element to the previous one. Although *E* might simply try each rule upon the first and second series elements, obtain a three-number series, and compare it to the given one, selecting the correct rule, he is forced into making a choice or judgment because none of the three alternative rules is fully correct. That is, no one rule will correctly reconstruct the series from the first element; but one will come nearest.

In the test, Way-Out Numbers, *E* is presented with a list of four ordered numbers and is instructed to choose either the first or last one on the basis of its being farther away from the remaining three numbers. In other words, *E* is to arrange the numbers on the dimension of numerical value and is to choose that extreme number whose value is farther from the other numbers' values.

EST—Evaluation of symbolic transformations. Three experimental tests were selected to measure the factor EST. The extreme difficulty of constructing reliable evaluative tests of symbolic transformations limited the choice of tests. The content of the three tests was concerned with changes from one form of symbol to another equivalent form, or changes in symbolic units to meet certain requirements. The transformations tests developed for this study used letters and words as stimuli. It appeared, during test construction, that numerical stimuli were not readily susceptible to transformations without the involvement of other products, such as relations or systems. The denotative inflexibility of numbers did not allow for equivalent forms of the same numerical value using two different symbols. This limitation had applied also in the study isolating the only other known symbolic-transformation fac-

tor, NST (Guilford, Merrifield, Christensen, & Frick, 1961). The tests loading on the NST factor all involved words as the stimuli.

A rather common transformation of words and letters is any code that allows their encoding. In almost all cases of symbol coding, the transformation of one set of symbols to its encoded set of symbols is a one-to-one mapping of the symbol set onto the code set. Such a one-to-one mapping is suitable for a test of sensitivity to errors in coding only when the test is speeded and the coding system is well known by the *E*s. This implies that the sensitivity to slight, but possibly important, miscodings is an evaluative process for the individual who functions well (is experienced) in the coding process.

Because no coding system is known with great generality within the population, and because it is inefficient and self-defeating to teach *E*s a complete coding system (memory factors might predominate), the EST test, Decoding, employs a simple and ambiguous code. Simplicity and ambiguity were introduced into the coding system for Decoding by employing a code for letters which does not map one-to-one onto the alphabet. The ambiguity of the code allows for words to be judged according to their ease of encoding or decoding. The change from an unambiguous code to an ambiguous one also changes the type of evaluation test involved. Whereas an unambiguous code and experience call for sensitivity, an ambiguous code calls for estimation; the code provides incomplete information, and *E* must estimate the complete information.

In the test, Decoding, *E* is presented with two words and is asked to choose which one, if coded, would be easier to decode unambiguously. *E* is also given the opportunity to judge both words as equal in difficulty of decoding.

Jumbled Words is the only test designed for EST that is in the sensitivity category. The *E* is given a stimulus word containing between five and seven letters and is to judge whether or not each of five alternative words is an accurate anagrammatic derivation from the given word. Jumbled Words is, therefore, similar in stimulus material to the ESU test, Derivations, which also uses anagram-type stimuli.

The third test designed for EST, Typing Errors, is similar to Decoding in the task involved and the stimuli used. The *E* is given an incorrectly typed word and is to choose from among alternatives the word that the incorrectly typed word would most likely be. The judgments are made on the basis of common typing errors due to the arrangement of the typewriter keyboard. A keyboard diagram is printed on each test page for *E*'s reference.

The estimation process involved in responding to Typing Errors is probably not dependent upon EST ability alone, however. It would seem that some figural ability would be involved in this test due to the spatial nature of the keyboard arrangement. Further, it might be expected that typing experience might enter into proficiency at the required task. However, the correlation of scores on Typing Errors and amount of typing experience was reported to be .03 (Hoepfner et al., 1964).

ESI—Evaluation of symbolic implications. Tests designed to measure the factor ESI employed all types of symbolic stimuli. For the evaluation process, implications are defined as the expectancies or probable relative values of the presented symbols (estimation), or possible symbolic interpretations of a unit or system (sensitivity to symbolic problems).

The test, Abbreviations, presented *E* with a shortened spelling of a common word, *E* to choose one of the three alternative words that the abbreviated word most likely implies. The meanings of the words are irrelevant to choosing an alternative, and the spelling of the alternatives is correct. The only task for *E* is to choose the most expected value for the abbreviation, a task of estimating. No observance of shorthand principles was exercised in test construction; the abbreviations were short and relatively unambiguous. Usually, but not always, this implied dropping vowels and unsounded consonants from the

keyed alternative. The *E* was warned, however, that sounding-out the abbreviation would not necessarily aid him in his choice. Hoepfner et al. (1964) found the correlation between scores on Abbreviations and experience with shorthand to be −.07.

Letter Problems is similar in format to Form Reasoning, a test of NSI. The evaluation form uses letters as the stimuli and asks *E* not to solve the equation, but to judge the difficulty or possibility of solving it, on the basis of provided rules. It was hypothesized that *E* would have to make his judgments based on foresight. The *E*'s judgments were of the three-category type; problems were easy to solve (straightforward), difficult to solve (involving manipulations), or impossible to solve due to inadequacies of the table of substitutions.

The third ESI test is named S Test. In this test, *E* is given a stimulus about which he is to find a problem to solve. The solution indicates the nature of the problem to which *E* was sensitive. The test might therefore measure *E*'s sensitivity to symbolic implications of unstructured problems. It should be noted, however, that this test is not congruent with the conception that evaluation is "sensitivity to *error*," for no error is judged. It is a test of *E*'s sensitivity to implications (as, indeed, it turned out) rather than a sensitivity to errors in implications.

Symbol Reasoning involves operations similar to the test, Logical Reasoning. The *E* is given two premises in the form of an equation involving inequalities such as $x < y = 3z$, and is asked to judge whether each of three conclusions (such as $x = 3z$) is true, false, or uncertain, on the basis of the given equation. The equation is a symbolic statement of the relationships between pairs of three unknowns, x, y, and z, in order. Each of the three conclusions to be evaluated involves one of the three possible pairings of unknowns.

Although it seemed reasonable to assume that conclusions involving adjacent unknowns, x and y, and y and z, might be relational, and conclusions involving the two extreme terms, x and z, would be more clearly implicational, pretesting analysis showed that the separately scored kinds of conclusions intercorrelated highly. For this reason, and for the reason that the numerical coefficients of the unknowns were not the same in the premise and the conclusions, it was decided that Symbol Reasoning should be a measure of ESI.

The Semantic Study

When a combination of letters is not only recognized as a group of symbolic entities, but also conveys meaning in the form of words, that meaning is semantic information. Semantic information constitutes the major content of verbal thinking. It can be transmitted through a number of media including words and sentences as well as pictures that imply verbal connotations.

Because of the richness of connotative meaning of words and the ambiguities inherent in our language structure, the concept of semantic evaluation may be comprehended easily at the conceptual level. However, in terms of test development, the existence of rich connotative meanings poses a difficult problem in insulating the processes of semantic evaluation from those of semantic cognition. Knowing or understanding the meaning of a word is a matter of semantic cognition. As it was stated previously, a test designed to measure an evaluative ability should pitch the evaluative aspect of the task at a level of difficulty that will ensure that individual differences in test scores reflect that source. On the other hand, the cognitive aspects of the task should be made so easy that they contribute a minimum to variance of the scores. If this approach is correct, we should be able to develop a test of evaluation from a test of cognition by emphasizing tasks that demand the evaluative operation and a minimum of cognitive operation. In fact, many tests in the category of semantic cognition, where all six factors have already been demonstrated, were used in this study as models for the development of tests of semantic evaluation. Such a procedure, if successful, en-

sures that the new factors are not merely the results of the difference in test formats.

Semantic reference factors. A major concern in this study is the separation of evaluative abilities from cognitive abilities. Six parallel semantic factors in the area of cognition were therefore included as reference factors. The reason for this concern regarding the separation of cognitive from parallel evaluative factors is that so many of the evaluation tests were constructed by analogy to cognition tests, and many of them resemble cognition tests except for rather subtle differences in emphasis. Even with the best of intentions, we could not expect that all cognitive variance would be eliminated from all such evaluation tests. The reference factors of semantic cognition and production are defined below, with mention of the tests used to represent them in this study.

CMU—Cognition of semantic units (verbal comprehension):

California Achievement Test—Reading Vocabulary—This vocabulary test is composed of 180 words in the four principal areas of the school curriculum—Mathematics, Science, Social Science, and General.

Verbal Comprehension—Select from alternatives the word that is similar in meaning to a given word.

Word Completion—Write the definitions or synonyms of given words.

CMC—Cognition of semantic classes:

Verbal Classification—Assign given words to one of two classes or to neither, each class defined by four other given words.

CMR—Cognition of semantic relations:

Verbal Analogies I—Discover the relation between two words and select the word that completes an analogy, the selection, as such, being quite easy.

CMS—Cognition of semantic systems (general reasoning):

California Achievement Test—Arithmetic Reasoning—This test consists of four sections—Number Concepts, Symbols and Rules, Numbers and Equations, and Problems.

Ship Destination Test—Find the distance from a ship to given points, considering the influences of several variables.

CMT—Cognition of semantic transformations (penetration):

Similarities—Write six ways in which common objects of a pair are alike.

CMI—Cognition of semantic implications (conceptual foresight):

Pertinent Questions—Write four questions, the answers to which would serve as a basis for making a decision in a conflict situation.

NMT—Convergent production of semantic transformations (semantic redefinition):

Picture Gestalt—Indicate which object in a photograph will serve a specified unconventional or uncommon purpose.

DMI—Divergent production of semantic implications (semantic elaboration):

Possible Jobs—Write as many as six different jobs which might be indicated by a pictured emblem.

Semantic evaluation factors. Of the five previously known factors mentioned in the introduction, four had been tentatively identified with semantic-evaluation cells of the structure-of-intellect model—"logical evaluation" with EMR, "experiential evaluation" with EMS, "judgment" with EMT, and "sensitivity to problems" with EMI. The coincidence of these four factors with structure-of-intellect abilities, however, was by no means regarded as firmly established. Consideration of relations between the four factors and the model led to the decision to develop new tests for all six of the hypothesized abilities. Some tests representing the previous factors or modifications of them were included in the new test battery in order to provide continuity with previous work.

Twenty-two tests were developed or extensively revised for this study. Eight pretest booklets of preliminary forms were administered to a number of classes at the University of Southern California. Two pilot studies were also conducted using high-school students who were expected to

be similar to those to whom the final battery was to be administered. These pretestings were designed to obtain various information concerning the tests and the examinee reactions to them. Information was obtained on such matters as the appropriate level of item difficulties, *E*'s comprehension of test instructions, test reliabilities, optimal time requirements for newly developed tests, and some preliminary intercorrelations. Extensive item analysis was conducted whenever the pretesting information indicated low reliabilities.

The hypothesized factors are defined below and the names of the tests used to represent them in the new battery are given. Further information regarding the tests, with sample items, may be found in Nihira, Guilford, Hoepfner, & Merrifield (1964).

EMU—Evaluation of semantic units: Units are relatively segregated or circumscribed items of information. A semantic unit may be a word, an object, an idea, or a verbalized concept, depending upon the nature of the information involved. In this study, the EMU factor is hypothesized to be the ability to evaluate the suitability or adequacy of a word or an object in terms of given criteria. Three tests were developed as measures of EMU.

In the test, Double Descriptions, *E* is to evaluate objects according to how they meet two stated criteria in the form of attributes. In each item, four alternative objects are to be judged, with the keyed object best.

The task in Synonyms is to evaluate the identity or degree of similarity of the meanings of words. This test is like most multiple-choice verbal-comprehension tests, except that *all* alternatives are synonyms of the given word. A choice must be made on the basis of subtle differences in meaning.

Word Substitution, the third test developed for EMU, asks *E* to evaluate a group of words in terms of their relative suitability in a given sentence. As in Synonyms, the decision regarding substitution of a word in a sentence is to be made among fine shadings of meaning, since any alternative could conceivably be chosen and would fit acceptably into the sentence.

EMC—Evaluation of semantic classes: Classes are recognized sets of units grouped by virtue of their common properties, In this study, the EMC factor is hypothesized as the ability to evaluate suitability or adequacy of a class grouping to represent a given word, object, or a set of objects.

From given alternatives in the test, Best Word Class, *E* is to choose the class name that best represents a given word or object. The alternative class names are all correct; choices are to be made on the basis of the criterion of how well the name covers the class properties of the object.

In the test, Best Word Pairs, *E* is to choose the pair that makes the best class from given pairs of words. The choice among word pairs, all pairs having properties in common, is to be made on the basis of the number and importance of shared properties.

The task in Class Name Selection is to choose the class name that best represents a set of words or objects from given alternatives. The alternative class names are all correct; choices are to be made in terms of aptness of the name, that is, which one describes the class most exactly.

EMR—Evaluation of semantic relations: Relations are recognized connections between units of information based upon variables that apply to them. In this study, the EMR factor is hypothesized to be the ability to evaluate relations between words or ideas. The factor called "logical evaluation," often identified in previous studies, was eventually assigned to the cell for EMR in the structure-of-intellect model. Therefore, Logical Reasoning, one of the tests that has consistently identified the logical-evaluation factor, was included in the present study to provide continuity with previous work.

In the test, Logical Reasoning, *E* is to choose the correct conclusion that can be drawn from two given premises. Only one of the alternatives in this syllogistic test is correct in each item. The incorrect ones are not obvious, however, since they repre-

sent common errors made in syllogistic reasoning.

In Best Trend Name, *E* is to select the word that best describes the order of four given words. All three alternative trend names refer to trends that at least partially describe the four words, but one describes a trend with greatest justification. This test is parallel to Seeing Trends, in which *E* names each trend, a test that probably measures both CMR and NMU.

The task in Matched Verbal Relations is to select the pair of words that represents the relationship most similar to the relationship given in the model pair of words. The difficulty is in the choice among alternatives, and not in the discovery of the relationship between the model word pair. The alternatives all have some plausibility in that they are related in some way with one another and with the model pair of words, but one pair exhibits a relationship most like that in the model word pair.

The fourth test, Verbal Analogies III, asks *E* to choose the alternative that is the best completion of the analogy, the relation between the first two words being fairly obvious. The keyed alternative has a greater similarity of relation to the completed analogy than do the other alternatives.

The *E* is to choose a word that is similar in meaning to each of two other given words in the test, Word Linkage. The word chosen must be related to the given words in two different ways. Only one alternative clearly conforms to both criteria; the remaining alternatives do not conform to one of them.

EMS—Evaluation of semantic systems: Systems are organized or structured aggregates of information, or complexes of interrelated parts. Present knowledge concerning the CMS and NMS factors suggests the diversity of characteristics of semantic systems. It seems that the semantic system can be a sentence, a complex of relationships among words, a problem, a sequence of events, or a common situation. For this reason, the tests developed for the EMS factor sampled a variety of problem items that could be expected to involve semantic systems.

Complete Thoughts is a new test in which *E* is to decide whether or not a given sentence expresses a complete thought. Sentences of the kind that characteristically confuse students as to whether they express complete thoughts were selected for this test. Completeness is the criterion.

From given alternatives in the test, Important Facts, *E* is to select the most important and the least important facts needed to solve a problem. All the alternative facts could possibly play a part in solving the problem, but one is most important under the given circumstances and one is least important.

Sentensense is like Complete Thoughts in that both present *E* with sentences. In Sentensense, *E* is to evaluate the internal consistency of the ideas or events expressed in each sentence. On the surface all sentences may appear to be meaningfully consistent, but some are not. This test was conceived as a verbal counterpart to the next one, Unlikely Things, which presents internal inconsistencies in pictorial form.

The task in Unlikely Things is to select from four given alternatives the two more unlikely things in sketches of a common situation. Judgments of unlikeliness in this test must be made on the basis of apparent violation of physical or conventional principles of varying degrees of possibility or on the basis of internal consistency. This test is a multiple-choice form of the test, Unusual Details, which had strongly helped to determine the factor of "experiential evaluation," later identified with the factor EMS.

Word Systems is a new test parallel to systems tests of figural matrices. *E* is to evaluate the internal consistency of a matrix of words arranged in terms of three meaningful rows and columns. None of the three word-matrix alternatives is completely consistent, but one is most consistent and one is least consistent.

EMT—Evaluation of semantic transformations: A transformation is defined as a change. In the semantic area, this usu-

ally means a change of interpretation or use of various objects, ideas, concepts, and other verbal-meaningful materials. In this study, EMT has been hypothesized as the ability to evaluate changes of interpretation of various objects and stories. The following tests have been adapted from tests for the NMT and DMT factors, emphasizing their evaluative aspects.

Product Choice is a test similar to the test Object Synthesis. The *E* is to select an object that can be made most adequately for a specified purpose by combining two given objects. The *E* must evaluate the adequacy of the unconventional uses of the common objects to choose the best answer.

In the test, Story Titles, *E* is to choose the best title that gives a new interpretation for a short story. One alternative title is always best on the basis of relevance to the story and provision of a new view or interpretation of the story.

From a set of alternatives, *E* is to select an object that can be used most adequately for a specified, unusual purpose in the test, Useful Changes. The *E* must judge which object, used unconventionally, would most adequately perform the given task. All the alternatives could be used to perform the task, but one is better on the basis of practicality and efficiency.

EMI—Evaluation of semantic implications. The factor called "sensitivity to problems," discussed in the introduction, had been assigned to the cell for EMI. For this reason, two of the tests used to measure the factor (Apparatus Test and Seeing Problems) were included in this battery as potential measures of EMI, in addition to the new tests developed parallel to tests of semantic implications in the areas of cognition and divergent production.

Apparatus Test asks *E* to suggest two improvements for each of a number of common appliances. Responses are scored on the basis of whether *E* senses the need for realistic and desirable improvements in the objects.

The next two tests were included as marker tests for the factor previously known as "judgment," which they had helped to define (Berger et al., 1957). Although "judgment" had been tentatively identified with factor EMT, the tests do not resemble very much the new tests developed for that factor. They were given a place among the EMI tests in this study, without serious expectation that they would cohere with that group in the new analysis.

In Commonsense Judgment I, *E* is to select the two best reasons why a proposed plan is faulty among the five given alternatives. All the reasons were designed to appear reasonable, but two are either more important or seem to be more apt. In the test, Commonsense Judgment II, *E* is to select the best method of demonstrating the truth of a given statement. All the alternative methods would demonstrate the statement's truth, but with varying degrees of success and physical possibility. The best method is successful, efficient, and possible to carry out.

The task in Seeing Problems is to list problems that might arise in connection with common objects. This test was planned to measure *E*'s sensitivity to consequences and other implications of the use of objects.

Sentence Selection is a new EMI test designed to measure *E*'s ability to evaluate extrapolations. The *E* is to select the statement that is most probably true, in view of given information. Choice of alternative conclusions is to be made on the basis of the conclusion's necessity. All conclusions could be true, but one is more fully determined by the given statement.

Word Extensions employs items containing logically necessary implications of words. The *E* is to select the name of an object or attribute that is always implied by a given word. The alternatives are all implied by the given word, but one is invariably implied whereas the others are implied only with some restriction.

Procedures

The Symbolic Study

The sample utilized in this study consisted of the entire senior-class student population of a Southern California high school.[3] Although 131

[3] The authors are indebted to the administra-

boys and 180 girls participated in the testing, the sample was later reduced to 86 boys and 139 girls, for whom complete test data for all experimental factor tests were available. The only criterion for exclusion of *E*s from the sample was incomplete data on these measures.

Age and IQ information was available for 219 and 199 students, respectively. Generalizing from such demographic data available for most of the sample utilized, the estimated mean age was 17.4 years. The estimated mean IQ, computed from combinations of scores obtained from the several IQ measures, which were variously administered between the eighth and eleventh grades, was 110.4. Although IQs ranged from 80 to 151, no students were deleted from the sample on the basis of extreme indexes of general intelligence.

The total sample of *E*s was tested in the mornings and afternoons of 2 consecutive days. Each testing session required approximately 2 hours. The tests of each factor were so arranged that order effects and fatigue effects would be approximately equal for all factors expected to be demonstrated.

The testing conditions under which the battery was administered were almost ideal, with one major exception. The days of the test administration, unfortunately, were only 4 and 5 days after the tragic death of President John F. Kennedy. Both administrators and school personnel were aware that after the day of national mourning, the preceding Monday, the students were still disturbed and restless. The effects of the national tragedy upon the results of this study are unknown.

Scoring criteria for the marker tests were developed from the scoring guides employed in previous studies. Scoring criteria for the newly developed experimental tests were based upon preliminary results of pretestings with university students in undergraduate psychology courses.

Frequency distributions were obtained for all part and total scores to determine whether the variables would meet the requirements of the Pearson *r* coefficient. A normalizing transformation was applied to those variables that were moderately skewed or exceedingly platykurtic. Extremely skewed or truncated variables were dichotomized near their medians. Descriptions of the frequency distributions and transformations for all the variables are listed in Table 1.

After it was ascertained that all the part-score data and total-score data, raw or scaled, met the requirements of the Pearson *r* or its approximations, reliability estimates were obtained from the raw scores of the tests. For all tests with two or more parts, Spearman-Brown reliability estimates were computed. Kuder-Richardson estimates of reliability were computed for all one-part tests that showed no evidence of speeding. Reliability estimates for one-part tests, wherein each item had a large possible range of scores, were computed by a formula suggested by Gulliksen (1950, p. 378).

Reliabilities of one-part speeded tests and of school measures could not be estimated. Their communalities were expected to approximate the necessary estimates of reliability.

The means and standard deviations of the variates are also listed in Table 1. In all but three cases, these descriptive statistics are based upon raw scores, before any transformations were applied.

Because the score matrix was incomplete and the data were differentially scaled, resulting in different kinds of correlation coefficients, the correlation matrix was obtained from a program that computes correlation coefficients between variables based upon the total number of individuals for whom scores are available. Most of the correlation coefficients in Table 2, therefore, are based upon the whole sample of 225 *E*s, but some are based upon the three variables for which not all *E*'s scores were available. These variables and the number of scores available are: Variable 52, 187 scores; Variable 53, 109 scores; and Variable 54, 107 scores. The attenuated samples for these variables taken independently, of course, result in further attenuation of sample size for the coefficients among them. The coefficient between Variables 52 and 54 was computed from a common sample of only 66 *E*s. This sample size was the smallest from which any coefficients were computed and was considerably smaller than the next smallest sample of 87 for the correlation between Variables 52 and 53.

Such variations in sample size upon which correlation coefficients are based introduce additional possibilities of error into the correlation matrix due to the necessary generalization that each coefficient in the matrix estimates equally accurately the actual intercorrelation between the variables. For each pair of variables, the coefficient based on the available *E*s is the best estimate, and since the *E*s in the reduced samples appeared to have been selected on irrelevant variables, that is, their attendance at the school when the tests were administered, any bias was thought to be negligible.

An additional consideration of the program employed is that it computes product-moment correlation coefficients upon any input data. This means that some coefficients are point-biserial *r*'s and some are phi coefficients. Standard corrections (Guilford, 1965, pp. 324, 353) were applied to each kind of coefficient to improve it as an estimate of the Pearson *r*. Thus, the coefficients reported in Table 2 are all Pearson *r* coefficients or estimates of the Pearson *r*.

The correlation matrix was submitted to an iterative communality-estimation program for estimates of communalities to be inserted into the principal diagonal for factor analysis. The iterated, stabilized communality estimates were put into the diagonal cells of the correlation matrix, and the matrix was submitted to a program that extracts principal-axes factors until the eigenvalues become negative, at which point extractions are

tion, staff, and students of Claremont High School for their splendid cooperation.

TABLE 1

MEANS, STANDARD DEVIATIONS, AND RELIABILITIES OF SYMBOLIC SCORES

Test	Mean	Standard deviation	Reliability[c]
1. Abbreviations–ESI01B	11.67	4.20	.26[d]
2. Best Letter Set–ESS03A	13.47	5.41	.56
3. Best Number Class–ESC01A	22.67[a]	6.74	.87
4. Best Number Pairs–ESC02A	17.32	5.99	.73
5. Camouflaged Words–NST01A	8.33	2.89	.74[d]
6. Circle Reasoning–CSS01D	6.84	2.72	.67[d]
7. Correct Letter Orders–ESS04A	16.04	8.17	.58
8. Correct Number Series–ESS05A	19.66	9.67	.74
9. Correct Spelling–ESU04A	36.98	11.49	.75
10. Decoding–EST01A	16.20	6.39	.74
11. Derivations–ESU08A	99.96	20.08	.81
12. Disemvowelled Words–CSU04B	11.53	4.21	.79[d]
13. Familiar Letter Combinations–ESU05A	14.23	6.59	.43[d]
14. Form Reasoning–NSI02C	17.49[a]	5.26	.96[d]
15. Identical Forms–EFU02A	38.12	7.30	.63[e]
16. Jumbled Words–EST03A	36.83[a]	10.86	.75
17. Letter Problems–ESI02A	14.95[b]	8.88	.88
18. Letter Triangle–CSS02B	5.68	2.85	.55[d]
19. Letter "U"–ESU06A	53.63	11.28	.84
20. Marking Speed Test	94.24[b]	18.12	.44[e]
21. Number Classification–CSC03B	11.41[b]	3.35	.72[d]
22. Number Grouping–DSC01B	14.10	4.42	.79
23. Number-Group Naming–CSC05A	10.17[a]	2.12	.76[d]
24. Numerical Operations–MSI01B	22.23	8.62	.78
25. Operations Sequence–NSS01B	12.27	5.29	.80
26. Perceptual Speed–EFU01A	48.01	9.21	.65[e]
27. Related Words I–ESR03A	14.00	5.41	.55
28. S Test–ESI04A	8.55	3.42	.69[d]
29. Seeing Trends II–CSR01B	8.16	3.25	.80[d]
30. Series Relations–ESS06A	9.93	7.22	.74
31. Sign Changes–NSI01A	17.49	5.05	.57
32. Sign Changes II–ESR01C	17.31[a]	3.64	.82
33. Similar Pairs–ESR04A	20.14[b]	7.40	.74
34. Sound Grouping–ESC04A	11.92	6.59	.74
35. Symbol Grouping–CSI01B	11.22	4.84	.84[f]
36. Symbol Identities–ESU07A	72.12	14.21	.90
37. Symbol Manipulation–ESR02C	21.40[a]	8.27	.74
38. Symbol Reasoning–ESI03A	17.95[b]	10.29	.78
39. Typing Errors–EST02A	9.44[b]	4.99	.49
40. Varied Symbols–DSC03B	10.92	4.50	.67
41. Way-Out Numbers–ESS07A	23.48[a]	6.55	.76
42. Word Changes–NSS02C	11.07[a]	4.47	.87[f]
43. Word Choice–ESC03A	13.05	5.81	.62
44. Word Combinations–CSU06A	10.05[b]	6.42	.72
45. Word Patterns–CSI03C	70.39	9.14	.75
46. Word Relations–CSR02B	10.76	4.64	.78
47. Word Transformation–NST02B	26.69	7.91	.83[f]
48. ITED General Vocabulary–CMU	20.35	5.45	.81[e]
49. PSAT Verbal–CMU	47.88	11.62	.95[e]
50. SCAT Verbal–CMU	304.82	14.33	.85[e]

[a] Total scores dichotomized at the medians for intercorrelations.

[b] Total scores C-scaled for intercorrelations.

[c] All estimates of reliability are Spearman-Brown corrections of correlation between parts unless noted.

[d] Kuder-Richardson estimate of reliability.

[e] Communality entered as reliability estimate.

[f] Reliability estimated through formula 21.21, in Gulliksen (1950).

stopped. This procedure resulted in the extraction of 32 real principal-axes factors. The first 18 factors accounted for 93.7% of the total variance of the 32-factor matrix and are presented in Table 3 as the unrotated factor matrix.

The principal axes were submitted to a program designed to rotate the loadings as closely as possible to a fixed target matrix of loadings (Cliff, 1964). The construction of the target matrix depended upon the intuitively inferred structure of the empirical matrix, simple structure, positive manifold, and the factor hypotheses. Successive adjustments of the target matrix effected considerable improvement in the empirical rotated matrix on all four criteria. The result of the final target-oriented rotation upon the principal-axes matrix is presented in Table 4 as the rotated factor matrix.

The Semantic Study

A high school in a newly developed industrial community in Los Angeles County cooperated in this study.[8] The final testing battery was administered in 2-hour sessions on each of 4 days, interspersed during a 2-week period. Approximately 400 eleventh-grade students participated in at least one of the testing sessions. The tests were organized in eight booklets, each of which required about 55 minutes. Rest periods were introduced between administrations of booklets.

An inspection of answer sheets indicated that a language IQ of at least 95 is necessary for adequate understanding of some of the test instructions. For this reason, the final sample was composed of students who completed all of the test booklets and who had both California Test of Mental Maturity language and nonlanguage IQs of 95 or above.

Reliabilities were estimated by correlating part scores and applying the Spearman-Brown formula. In an effort to increase reliabilities, several tests were item analyzed when their reliabilities were found to be lower than .50. After the item analyses, the internal-consistency reliabilities were estimated for these tests. Test reliabilities are reported in Table 5 along with means and standard deviations.

In this study, two separately timed parts of each of six reference tests were employed to define the pertinent reference factors. The reliabilities of these reference tests are indicated by correlations between the two separately timed parts, not by applying the Spearman-Brown formula.

The use of alternate forms of certain factor tests should increase the chances of appropriate location of the axes for the factors that those tests represent, but there is the disadvantage that loadings in those tests may involve specific as well as common-factor variance. Loadings for the factors in other tests should presumably not be affected. Since it is this reference-factor involvement in the experimental test that is to be accounted for, the use of alternate forms of marker tests can thus be justified.

Frequency distributions were inspected for irregularity. All test scores were accepted as appropriate for Pearson product-moment correlation coefficients. The correlation matrix is given in Table 6.

In extracting principal-axes factors, a computer program for iterative factor analysis was applied. The iterative factor-analysis program extracts a specified number of principal-axes factors using estimated communality values. In the first iteration, the program reextracts principal-axes factors using the computed communalities obtained from the result of the first extraction. Starting with the highest correlation in each column as the initial estimate of communalities, the iterative factor-analysis program iterated the principal-axis extraction process nine times until the communalities became stabilized. The communalities changed less than .01 between the eighth and ninth cycles of iteration.

The extraction of principal-axes factors is equivalent to choosing a set of factors in decreasing order of their contribution to the total variance of the matrix. This principle provides a rough numerical guide as to the number of factors to be retained for rotational solution. Since the observed correlations are subject to error of estimate, it was decided to retain the largest number of extracted factors whose cumulative contribution accounted for less than 95% of the total variance. According to this criterion, 14 principal axes were retained and iterated and are presented in Table 7.

Graphic orthogonal rotations were used to locate the new reference axes. During the initial phase of the rotation processes, the objective criterion of simple structure was the primary consideration. The secondary aim was to spread the variance of factor loadings as equally as possible among the factors. After each axis had been rotated at least once, and the reference factors defined by the marker tests began to emerge, positive manifold and psychological meaningfulness became the important criteria. The final rotations consisted of computer adjustments, employing Cliff's (1964) procedure, aimed at the improvement of positive manifold and simple structure. The rotated factor matrix is presented in Table 8.

Results of the Factor Analyses

The interpretations of rotated factors rest principally upon tests with factor loadings of .30 or greater. The names of the tests and the factor for which each test was designed are listed preceding the discussion for each obtained factor. A test is listed if it has a loading as large as .30. If a test

[8] The authors wish to thank the administration, staff, and students of the La Puente School District for their excellent cooperation.

34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50
33	29	24	25	33	13	03	25	48	27	29	25	32	31	36	30	27
41	38	24	38	42	29	11	24	39	34	35	20	37	36	41	25	38
48	54	30	68	62	29	−01	59	49	39	35	33	48	54	54	39	30
49	44	34	48	49	37	09	46	47	41	36	30	54	40	56	40	41
29	06	09	19	13	14	12	08	19	02	23	25	20	33	11	39	17
29	30	08	33	36	16	07	14	34	34	30	28	29	33	30	23	11
43	41	30	40	56	28	12	48	49	36	42	30	49	41	65	51	47
50	56	41	50	60	28	17	53	56	46	45	36	57	42	56	52	47
51	22	42	26	33	20	09	23	29	22	40	26	35	49	49	47	44
42	40	30	38	44	30	16	29	44	41	35	26	40	39	43	46	29
23	29	47	31	30	20	26	25	29	32	23	29	39	33	16	14	08
42	13	20	15	15	15	27	13	34	12	37	27	25	46	20	20	30
18	18	09	16	18	11	02	10	20	14	23	18	27	23	16	13	07
23	27	37	30	22	32	09	35	30	29	27	23	27	11	20	15	32
23	26	50	19	28	15	19	30	33	17	22	27	39	24	25	27	15
35	39	49	38	41	29	19	36	58	41	45	38	59	44	36	56	34
35	38	32	29	39	23	07	26	34	40	31	31	40	23	45	28	21
28	36	21	45	42	26	11	33	53	30	28	27	43	32	34	25	23
13	26	51	14	18	17	12	20	32	15	15	24	14	19	11	11	−05
−01	01	27	−11	−02	01	−02	11	05	00	02	08	03	09	−10	−03	−19
38	45	33	46	43	33	08	35	42	37	29	31	43	42	36	26	31
42	48	40	53	51	34	13	41	51	47	39	40	47	44	48	42	37
45	45	31	53	53	28	11	49	66	44	44	43	49	40	52	31	40
11	21	41	06	17	13	13	41	29	17	20	15	25	27	06	04	06
47	45	37	50	55	40	21	38	47	39	33	41	49	35	37	28	37
19	37	32	19	28	15	12	19	23	23	18	24	34	19	22	26	24
40	48	30	49	41	32	14	33	44	40	25	19	59	27	43	26	33
08	18	17	00	11	10	22	10	04	11	08	16	09	09	16	−09	01
38	31	22	44	47	22	14	35	49	31	36	32	52	36	54	55	46
39	47	24	45	51	31	11	48	54	32	35	27	48	38	42	32	44
17	33	51	16	29	21	16	44	40	20	17	27	34	26	17	12	02
41	42	25	39	40	23	18	39	43	33	26	26	41	33	38	19	30
40	47	34	51	48	35	09	34	49	41	37	26	55	33	50	33	41
	34	26	40	44	25	15	26	43	30	44	33	47	47	60	52	56
34		27	43	44	26	09	36	45	39	32	34	43	27	44	23	19
26	27		14	31	19	18	33	38	26	20	26	35	24	35	26	11
40	43	14		56	30	03	30	48	44	26	39	45	36	43	44	50
44	44	31	56		32	11	49	40	38	36	31	54	33	57	52	42
25	26	19	30	32		11	20	23	19	23	17	31	26	27	24	19
15	09	18	03	11	11		03	20	15	13	21	26	20	10	00	05
26	36	33	30	49	20	03		49	25	25	14	40	24	35	30	30
43	45	38	48	40	23	20	49		38	40	34	58	49	43	52	34
30	39	26	44	38	19	15	25	38		32	36	44	27	37	24	35
44	32	20	26	36	23	13	25	40	32		31	43	47	39	40	32
33	34	26	39	31	17	21	14	34	36	31		34	41	30	27	12
47	43	35	45	54	31	26	40	58	44	43	34		40	46	32	31
47	27	24	36	33	26	20	24	49	27	47	41	40		43	44	45
60	44	35	43	57	27	10	35	43	37	39	30	46	43		71	63
52	23	26	44	52	24	00	30	52	24	40	27	32	44	71		66
56	19	11	50	42	19	05	30	34	35	32	12	31	45	63	66	

TABLE 3

Unrotated Symbolic Factor Matrix

Variable	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	h²
1.	50	06	04	−08	−06	−10	−11	−18	−20	24	07	13	06	01	17	−05	00	−11	49
2.	60	02	−12	02	−03	04	22	−05	−03	−00	15	−04	32	−11	22	−01	−02	−04	62
3.	77	−04	−23	33	03	−02	12	−27	01	−03	−11	12	−09	−16	−04	10	06	−01	92
4.	74	11	−18	12	02	−10	04	03	−12	10	−05	−07	12	18	−18	−06	13	02	75
5.	29	04	39	20	−16	−17	−06	−05	19	−13	00	−03	27	03	−11	05	00	−01	48
6.	42	12	02	−16	−28	−04	−14	−17	−11	03	−03	−05	−16	−10	−01	−15	−08	−08	43
7.	69	18	−01	09	16	−04	−11	15	−07	−01	−10	−04	05	−08	06	10	−05	−14	64
8.	80	00	−06	07	08	−09	−06	06	01	00	−05	−04	−04	−06	01	−05	09	09	70
9.	57	07	38	09	07	15	22	−07	−24	05	−11	04	−12	07	−14	−02	02	−05	68
10.	63	05	−05	14	−05	−14	02	−01	−09	−02	−14	03	11	09	14	−10	−14	06	55
11.	49	−38	09	−27	−04	05	06	−08	10	−06	−20	12	−00	04	09	12	−07	−08	58
12.	38	−04	49	05	−29	22	10	07	02	05	14	08	−10	−08	−06	−03	02	−03	59
13.	30	−06	06	−15	−09	05	−01	−27	−08	22	−05	−12	09	−06	−10	08	−01	10	31
14.	44	−20	−04	−32	23	31	19	−09	−03	04	12	−31	−04	03	−06	03	−04	−09	66
15.	46	−32	12	−19	11	−36	12	−02	22	16	11	08	−03	−04	−07	−07	−05	−02	63
16.	71	−15	21	−19	02	16	−28	06	−01	−09	−20	−03	−05	08	−02	−14	13	17	80
17.	55	−00	−07	−20	−00	−12	−09	08	−30	−11	15	−05	11	−07	−00	12	02	11	55
18.	56	01	−18	−10	−14	11	−21	03	11	01	18	05	06	−04	−02	05	19	02	52
19.	37	−52	12	06	01	−08	−07	−07	−14	−20	30	09	−03	01	04	03	06	01	61
20.	07	−51	18	13	09	−05	−11	−17	−04	−07	02	−15	01	−08	−02	01	−12	09	42
21.	65	−07	−16	12	−07	−01	18	−09	−05	04	−11	06	05	10	04	−07	18	−09	59
22.	74	−01	−08	08	−03	−15	01	−07	−06	−10	−08	−12	−02	12	02	−13	−06	−08	69
23.	74	04	−15	19	−17	−13	02	−01	08	21	11	−17	−05	20	−05	07	−06	00	80
24.	38	−53	06	28	18	29	−02	08	01	03	−14	−04	04	−07	14	10	09	07	70
25.	72	−14	−12	12	−13	10	15	06	06	−15	14	−02	02	−00	−04	−08	−06	08	68
26.	43	−24	01	−38	16	−24	20	−12	28	12	−06	−03	01	−10	03	−11	09	07	66
27.	62	02	−18	−23	−01	18	10	02	−01	−01	−10	23	12	06	−03	03	−10	03	61
28.	17	−12	−01	−01	−12	−25	16	39	−02	−07	−05	−10	−07	−08	02	−06	16	−07	36
29.	61	24	05	−11	13	−05	−20	11	04	01	−05	−15	15	00	−02	13	04	−12	60
30.	65	04	−08	19	01	16	−13	12	05	−07	04	−02	−05	−11	01	−23	−14	−03	62
31.	52	−54	−04	21	10	03	−04	−01	03	01	01	01	09	04	−00	−02	−06	−04	64
32.	57	−02	−18	19	−13	−05	08	15	05	14	03	04	−09	08	−03	09	00	21	53
33.	65	13	−11	−28	01	18	−07	01	−08	−04	08	13	05	−03	−07	−03	10	−04	62
34.	65	26	24	06	−07	02	21	02	−07	05	10	09	01	−06	−06	07	−12	14	67
35.	62	−06	−27	−08	−07	−07	00	02	−05	−02	−01	05	−09	−18	04	−08	04	06	54
36.	52	−46	17	−11	25	−10	04	12	−15	−06	05	14	−11	19	−05	03	−00	−07	72
37.	65	26	−24	−03	−11	09	02	−24	27	−28	04	04	−13	11	−02	13	02	−04	83
38.	70	15	−15	02	09	−08	−03	−00	04	−13	−02	05	−07	−17	−08	02	−12	−05	63
39.	42	−01	−07	−04	−04	11	11	03	−01	−19	03	−05	19	02	−17	−15	−00	−09	35
40.	22	−15	17	−12	−24	−01	08	37	13	00	−06	03	01	06	06	05	−10	−06	36
41.	57	−12	−19	21	29	09	−14	10	11	17	01	−04	−13	−09	−11	02	−03	−10	63
42.	73	−02	04	02	−07	08	−35	02	13	21	19	08	−05	14	12	−04	−02	−06	82
43.	56	04	−16	−18	−11	01	01	02	−09	−09	−09	−13	−14	10	22	05	−08	10	52
44.	55	10	23	−03	−12	06	−04	−00	−10	09	−03	−17	03	−16	01	−02	−01	02	46
45.	49	−08	13	−07	−28	−15	−03	−05	02	−14	−00	−19	−15	02	−00	16	03	−05	48
46.	71	−01	−04	−19	−08	07	−11	14	09	10	−15	09	10	−03	−16	06	−11	06	69
47.	62	07	35	16	−21	06	04	−10	05	04	−11	02	−01	−08	10	03	12	−09	65
48.	69	39	07	−02	23	−17	08	12	−19	00	05	08	−08	−09	01	13	−01	−03	81
49.	61	42	40	02	37	−16	−18	−15	11	−15	−01	04	02	06	02	−05	00	08	98
50.	56	51	19	00	28	18	25	07	20	02	13	−08	−05	08	18	−06	04	07	90

Note.—Decimal points omitted.

TABLE 4

Rotated Symbolic Factor Matrix

Test Name	CSU	CSC	CSR	CSS	CSI	CMU	MSI	DSC	NSS	NST	NSI	EFU	ESU	ESC	ESR	ESS	EST	ESI	h^2
1. Abbreviations ESI	17	17	21	19	00	10	06	00	21	06	−09	14	09	13	−04	07	10	47	49
2. Best Letter Set ESS	01	19	−01	−04	17	20	14	06	15	26	22	07	−07	14	20	02	19	53	66
3. Best Number Class ESC	−05	32	05	12	24	10	29	−17	10	31	−02	13	15	50	27	32	04	23	93
4. Best Number Pairs ESC	−08	54	26	23	12	20	06	−03	11	18	14	06	03	27	05	15	29	13	76
5. Camouflaged Words NST	04	−01	06	05	00	12	−08	07	12	62	−03	02	14	00	−05	01	16	−04	48
6. Circle Reasoning CSS	27	01	22	30	24	−05	−12	−07	19	03	−08	08	01	15	08	11	17	17	44
7. Correct Letter Orders ESS	03	21	29	05	15	32	11	10	03	18	01	00	03	18	10	43	22	26	64
8. Correct Number Series ESS	−05	22	18	19	25	27	19	07	20	16	09	17	11	28	08	31	27	16	69
9. Correct Spelling ESU	47	28	13	01	02	28	13	−08	06	15	07	01	29	29	01	11	23	−01	68
10. Decoding EST	−06	24	12	00	11	12	02	01	22	18	−05	05	08	29	05	19	37	27	53
11. Derivations ESU	10	−02	22	−10	08	−11	20	16	11	09	12	31	34	14	29	03	21	16	57
12. Disemvowelled Words CSU	55	02	−07	11	07	10	07	24	24	28	06	00	16	11	04	03	08	−05	59
13. Familiar Letter Combinations ESU	18	00	26	14	−03	−07	05	−08	02	14	20	15	−01	26	−02	−07	04	16	32
14. Form Reasoning NSI	22	10	26	−07	14	13	13	−10	08	−13	59	16	09	06	23	04	05	12	67
15. Identical Forms EFU	00	14	03	03	12	07	−02	13	14	15	15	63	27	03	−04	10	06	14	63
16. Jumbled Words EST	11	−06	35	28	08	16	27	12	29	04	17	12	27	17	10	17	48	−05	84
17. Letter Problems ESI	−04	08	21	23	24	18	−04	17	−03	03	22	−03	18	16	00	08	22	38	54
18. Letter Triangle CSS	−06	10	19	39	13	11	14	21	24	12	14	05	−01	09	26	11	04	20	50
19. Letter U ESU	−04	01	−12	10	19	−06	18	09	23	12	24	04	56	01	−05	06	03	23	61
20. Marking Speed	−03	−20	−05	−09	05	−22	16	−12	09	17	27	08	33	04	−21	12	02	03	43
21. Number Classification CSC	00	43	11	07	19	07	27	−04	18	17	01	12	09	24	20	03	22	21	59
22. Number Grouping DSC	−03	31	23	09	29	12	02	−10	26	17	07	09	17	21	11	22	34	19	65
23. Number-Group Naming CSC	−03	47	28	12	22	07	−06	09	38	25	14	09	00	35	03	21	04	13	80
24. Numerical Operations MSI	−01	01	−02	−19	00	−07	61	11	15	14	28	00	21	16	01	26	10	05	70
25. Operations Sequence NSS	00	26	−06	11	27	09	08	11	30	21	29	03	13	28	27	19	24	14	68
26. Perceptual Speed EFU	−04	01	13	−01	18	14	12	02	07	05	18	69	09	07	11	−07	15	12	68
27. Related Words I ESR	04	21	16	11	−06	11	06	16	06	−01	15	15	08	29	43	09	28	25	60
28. S Test ESI	−02	19	−05	−05	41	01	06	27	−07	03	−03	11	01	−12	−09	04	19	−04	36
29. Seeing Trends II CSR	01	12	46	14	09	35	02	12	03	22	09	06	−02	01	10	26	19	18	60
30. Series Relations ESS	07	14	00	18	16	17	11	02	31	11	12	−02	−01	14	18	48	27	12	62

31. Sign Changes NSI	−10	19	−01	−04	04	−15	33	05	24	21	30	15	32	10	03	28	15	17	64
32. Sign Changes II ESR	−09	35	06	09	15	06	06	25	23	10	08	07	01	43	06	19	07	04	54
33. Similar Pairs ESR	12	16	23	36	07	26	08	11	09	−03	16	08	08	12	35	11	22	26	63
34. Sound Grouping ESC	30	20	04	08	07	38	−11	15	11	27	07	04	08	41	08	13	16	16	66
35. Symbol Grouping CSI	−05	16	07	24	33	06	13	08	11	−02	06	19	03	25	19	21	20	27	52
36. Symbol Identities ESU	04	23	10	−04	07	08	17	16	12	−06	23	23	62	03	−03	17	19	10	72
37. Symbol Manipulation ESR	−11	15	25	19	25	24	−07	−08	26	19	−01	01	08	31	59	09	01	04	83
38. Symbol Reasoning ESI	−03	17	14	18	24	25	−02	00	04	16	04	14	08	24	27	41	17	20	62
39. Typing Errors EST	04	21	01	14	12	09	01	−04	05	17	25	−01	05	−01	26	05	30	08	35
40. Varied Symbols DSC	15	06	05	−13	12	−07	−03	46	11	09	03	09	06	−05	08	05	19	−04	36
41. Way-Out Numbers ESS	−06	27	14	10	05	12	27	01	14	05	20	16	04	14	08	57	−03	06	64
42. Word Changes NSS	08	16	29	29	−03	15	11	22	56	11	06	12	12	11	09	32	07	23	82
43. Word Choice ESC	−01	07	31	00	32	09	−01	10	22	−12	07	00	06	31	19	07	25	20	53
44. Word Combinations CSU	32	02	22	14	18	17	07	07	11	23	11	03	−03	19	−04	16	23	17	47
45. Word Patterns CSI	12	03	29	08	42	−02	−05	12	17	25	04	04	22	16	07	03	09	04	48
46. Word Relations CSR	08	15	31	23	−01	06	04	28	08	15	17	23	03	26	25	27	30	12	68
47. Word Transformations NST	35	09	16	06	16	19	22	07	24	43	−13	06	10	22	09	10	17	11	66
48. ITED Verbal CMU	12	27	22	10	18	57	−07	10	−08	07	−06	07	11	26	04	31	14	28	81
49. PSAT Verbal CMU	03	−05	28	14	−06	72	−08	−15	18	28	−11	14	23	13	00	24	24	03	98
50. SCAT Verbal CMU	19	15	12	−10	04	79	−01	01	26	06	06	07	14	19	21	10	10	02	91

Note.—Decimal points omitted.

TABLE 5

Means, Standard Deviations, and Reliabilities of Semantic Scores

Test	Mean	Standard deviation	Reliability[a]
51. Reading Vocabulary (CAT)	103.8	24.4	.88[b]
52. Arithmetic Reasoning (CAT)	101.1	27.4	.89[b]
53. Apparatus Test–EMI05C	9.7	4.3	.71
54. Best Trend Name–EMR01A	11.4	3.6	.73
55. Best Word Class–EMC01A	20.7	3.4	.61[c]
56. Best Word Pairs–EMC02A	18.8	4.7	.73
57. Class Name Selection–EMC03A	21.4	3.7	.63
58. Commonsense Judgment I–EMI04A	5.6	1.6	.52[d]
59. Commonsense Judgment II–EMI04B	4.2	1.5	.26[d]
60. Complete Thoughts–EMS01A	39.0	6.5	.77
61. Double Descriptions–EMU01A	35.2	4.2	.61
62. Important Facts–EMS02A	11.8	3.0	.49
63. Logical Reasoning, Form A-2	12.4	4.5	.82
64. Matched Verbal Relations–EMR02A	14.7	4.8	.65
65. Pertinent Questions, Form A-J (Part I)	9.2	2.6	.65[e]
66. Pertinent Questions, Form A-J (Part II)	11.0	2.6	.65[e]
67. Picture Gestalt–NMT03B	20.0	2.4	.64
68. Possible Jobs–DMI03A (Part I)	8.0	3.0	.67[e]
69. Possible Jobs–DMI03A (Part II)	7.7	3.5	.67[e]
70. Product Choice–EMT01A	34.6	5.6	.57[f]
71. Seeing Problems–EMI01B	14.7	5.2	.73
72. Sentence Selection–EMI02A	9.6	2.7	.46
73. Sentensense–EMS03A	15.0	2.3	.50[c]
74. Ship Destination Test, Form A-2	22.5	9.9	.67[d]
75. Similarities–CMT02B (Part I)	13.5	3.2	.54[e]
76. Similarities–CMT02B (Part II)	10.2	3.3	.54[e]
77. Story Titles–EMT02A	21.7	7.0	.67
78. Synonyms–EMU02A	22.2	3.9	.39
79. Unlikely Things–EMS04A	23.6	2.8	.54
80. Useful Changes–EMT03A	16.7	3.7	.53
81. Verbal Analogies I–CMR01A (Part I)	5.2	1.9	.38[e]
82. Verbal Analogies I–CMR01A (Part II)	9.2	2.6	.38[e]
83. Verbal Analogies III–EMR04A	8.1	3.1	.60
84. Verbal Classification–CMC02A (Part I)	25.5	8.1	.64[e]
85. Verbal Classification–CMC02A (Part II)	30.4	10.1	.64[e]
86. Verbal Comprehension, Form B	15.4	3.7	.68
87. Word Completion–CMU01A	9.2	3.6	.75
88. Word Extensions–EMI03A	17.5	4.3	.63
89. Word Linkage–EMR03A	22.0	4.6	.74
90. Word Substitution–EMU03A	16.1	3.5	.51
91. Word Systems–EMS05A	11.6	3.3	.28[f]

[a] All reliability estimates are based upon alternate-form correlations extended by the Spearman-Brown formula, except as noted.

[b] Reported by the test publisher.

[c] Kuder-Richardson formula 20.

[d] Communality as the lower bound estimate of reliability.

[e] Correlation between the two parts of the same reference test.

[f] Hoyt's method for estimating reliability.

has loadings of .30 or greater on other factors in the solution, those loadings and their factors are mentioned in parentheses.

Emphasis is placed upon the discussion of newly isolated factors of evaluation. The factors of cognition and other reference factors will be discussed to the extent that this helps to understand the nature of evaluative factors, or throws new light on the reference factors and their tests. Extensive discussion of all reference factors in this study has been given elsewhere in the series of *Reports from the Psychological Laboratory* at the University of Southern California.

78	79	80	81	82	83	84	85	86	87	88	89	90	91
28	21	22	45	37	37	34	39	52	54	34	37	51	18
28	22	19	49	43	47	43	46	39	41	38	32	42	29
11	19	12	17	17	11	20	11	20	26	16	18	20	14
34	30	23	46	44	55	38	43	44	46	48	25	47	34
29	20	29	41	36	37	42	42	35	37	32	41	28	22
27	31	17	43	44	37	43	48	29	34	44	30	31	22
37	28	29	39	41	39	47	42	38	51	46	31	43	25
30	37	23	37	34	32	39	37	31	37	34	28	38	16
15	11	08	12	14	05	14	01	18	19	02	21	12	17
47	32	27	37	37	41	42	39	45	41	43	26	42	16
26	29	17	21	25	22	30	24	21	22	23	25	28	19
30	25	18	31	32	36	37	39	28	31	39	24	29	13
40	34	22	45	49	47	42	42	42	41	47	28	40	28
42	40	29	56	58	62	51	54	47	51	57	39	55	33
17	15	06	00	04	−02	19	09	23	26	08	04	22	15
30	24	12	22	16	20	35	29	30	36	30	11	31	13
19	19	16	13	21	16	28	24	21	29	20	22	27	13
24	30	07	27	29	30	31	29	26	34	27	22	29	18
29	37	14	31	32	40	36	32	36	43	41	17	38	17
21	32	11	16	30	27	25	22	26	22	21	31	19	18
23	09	00	10	22	09	18	22	25	30	14	17	20	06
36	36	14	38	41	44	29	41	39	40	47	32	42	25
31	30	11	32	27	31	31	26	39	39	38	27	33	23
21	36	17	41	39	43	36	37	32	30	39	20	33	37
21	15	15	24	18	23	23	26	31	24	25	21	24	13
23	25	07	31	33	26	29	33	28	32	19	23	27	25
36	32	22	44	40	35	35	34	43	52	41	27	41	33
	29	29	29	35	28	30	34	48	40	38	32	43	17
29		17	26	37	32	35	32	30	37	40	31	35	33
29	17		23	15	27	14	11	40	34	26	14	33	−01
29	26	23		38	52	35	44	37	42	37	26	34	25
35	37	15	38		44	40	43	37	40	49	43	34	30
28	32	27	52	44		43	37	36	44	51	24	38	22
30	35	14	35	40	43		64	25	39	47	27	32	23
34	32	11	44	43	37	64		33	34	45	31	41	22
48	30	40	37	37	36	25	33		65	39	23	57	19
40	37	34	42	40	44	39	34	65		45	28	60	29
38	40	26	37	49	51	47	45	39	45		26	48	23
32	31	14	26	43	24	27	31	23	28	26		26	18
43	35	33	34	34	38	32	41	57	60	48	26		17
17	33	−01	25	30	22	23	22	19	29	23	18	17	

TABLE 7

Unrotated Semantic Factor Matrix

Variable	A	B	C	D	E	F	G	H	I	J	K	L	M	N	h²
51.	64	05	−35	12	36	20	15	−16	−01	−12	−06	−21	−09	04	84
52.	66	12	−09	08	41	20	07	−06	−12	−01	−07	−06	01	00	70
53.	36	−35	08	08	−00	09	05	−03	−15	23	−02	−00	08	−00	36
54.	69	13	−01	−07	07	14	−20	01	−04	03	−03	11	09	−03	59
55.	58	12	−06	29	13	−20	15	08	−04	25	−02	12	02	−00	60
56.	58	13	11	04	11	−08	02	06	−07	−03	03	01	04	−06	40
57.	66	10	−07	04	−02	−20	14	04	−04	07	00	13	−03	23	59
58.	56	09	02	03	−10	−22	05	03	−04	11	−06	−14	−15	09	45
59.	20	−17	−10	25	01	14	18	18	−05	09	−01	08	−01	−14	26
60.	65	10	−07	−07	−13	−20	02	00	−07	15	−17	−08	11	08	58
61.	43	−05	29	25	−33	13	02	−19	−20	−18	−05	03	−14	04	59
62.	54	08	06	−03	01	−21	07	−07	−05	07	−09	−02	−12	10	40
63.	67	15	11	02	−04	−01	−13	−11	−00	02	−05	06	14	12	56
64.	80	24	09	−04	−05	04	−09	−06	06	−06	−04	09	−00	−06	74
65.	29	−71	−12	11	09	−08	−09	20	−11	−00	−00	00	10	02	70
66.	47	−58	−04	−02	06	−18	−16	17	−11	−03	−04	−04	−03	−17	70
67.	40	−31	11	02	04	16	12	−17	−22	−09	13	03	06	07	42
68.	50	−34	14	−56	12	09	26	−10	04	17	03	−01	−00	04	83
69.	59	−22	04	−41	05	05	10	−05	−01	03	03	06	−06	−04	59
70.	40	−07	20	08	−23	09	22	−10	01	01	04	02	−08	−03	34
71.	33	−40	03	04	−08	03	12	−04	13	−26	−15	26	01	02	47
72.	62	18	04	−17	−08	07	−02	03	−00	04	−22	−10	18	04	55
73.	54	−07	01	−06	−21	06	−01	11	04	−07	−34	−05	−17	−06	52
74.	59	16	23	09	01	28	−22	04	−26	02	−04	−03	07	02	64
75.	41	−32	11	11	−04	−05	−26	−25	12	09	−06	−13	02	−15	49
76.	49	−39	20	21	11	−09	−26	−21	34	12	12	−03	−05	16	78
77.	62	02	−15	−01	01	05	−07	20	16	−05	−06	06	−12	−01	50
78.	55	−01	−16	01	−23	−12	05	01	08	−07	−03	−00	17	−02	44
79.	52	00	17	−08	−25	07	05	24	03	03	25	−28	−07	−01	58
80.	36	09	−34	09	−22	−09	−04	−16	−17	16	24	06	−02	−09	47
81.	62	20	−02	03	17	03	−14	−03	07	12	−05	05	−17	−10	54
82.	64	17	14	05	−00	08	10	04	21	−03	09	10	16	−04	57
83.	64	25	03	−14	04	03	−14	−07	00	10	07	15	−11	−17	60
84.	63	02	20	03	15	−25	08	11	−15	−14	11	04	−09	−03	61
85.	65	13	20	02	25	−32	02	01	04	−32	08	−09	06	−02	77
86.	65	−06	−43	03	−17	08	−08	−07	08	−06	02	−03	05	01	67
87.	70	−09	−35	−04	−09	09	01	08	08	−05	13	08	−11	10	70
88.	66	17	01	−18	−08	−12	−09	09	−04	−09	07	02	08	−12	57
89.	47	06	09	26	−06	05	39	−03	22	07	−01	−11	11	−18	57
90.	65	−01	−30	−06	−11	03	−06	−06	−04	−17	12	−11	06	07	60
91.	40	02	19	11	03	30	−11	37	12	07	08	−00	−01	16	50

Note.—Decimal points omitted.

Interpretation of the Symbolic Reference Factors

CSU—Cognition of symbolic units:

12. Disemvowelled Words (CSU) .55
9. Correct Spelling (ESU) .47
47. Word Transformation (NST) .35 (.43 NST)
44. Word Combinations (CSU) .32
34. Sound Grouping (ESC) .30 (.41 ESC; .38 CMU)

Disemvowelled Words once again leads

TABLE 8

ROTATED SEMANTIC FACTOR MATRIX

Test name	CMU	CMC	CMR	CMS	CMT	CMI	DMI	NMT	EMU	EMC	EMR	EMS	EMT	EMI	h²
51. Reading Vocabulary CMU	70	19	11	47	08	−03	15	−03	03	17	10	−04	06	−02	84
52. Arithmetic Reasoning CMS	40	24	14	57	02	01	15	07	−02	19	26	02	03	09	70
53. Apparatus Test EMI	01	−06	10	26	18	34	19	14	09	13	01	06	20	10	36
54. Best Trend Name EMR	29	12	07	28	05	07	07	19	04	08	48	14	10	29	58
55. Best Word Class EMC	16	19	23	22	05	12	−06	−03	−02	54	20	04	22	18	60
56. Best Word Pairs EMC	16	34	15	21	02	05	06	08	06	17	25	12	13	23	40
57. Class Name Selection EMC	31	21	07	05	00	07	08	16	10	50	17	11	12	29	58
58. Commonsense Judgment I EMI	21	21	−01	07	11	01	09	−09	21	33	11	14	18	33	45
59. Commonsense Judgment II EMI	11	−05	22	18	−08	28	−04	−05	07	14	01	10	15	−13	26
60. Complete Thoughts EMS	27	11	07	10	04	07	10	03	12	28	17	03	20	55	58
61. Double Descriptions EMU	05	14	05	16	13	06	−07	27	64	05	09	08	15	07	60
62. Important Facts EMS	17	24	00	09	10	01	16	01	18	33	17	01	09	28	39
63. Logical Reasoning EMR	23	16	12	18	17	−02	02	27	13	18	32	09	08	40	55
64. Matched Verbal Relations EMR	32	26	18	16	08	−05	08	16	23	14	50	13	12	34	74
65. Pertinent Questions (Part I) CMI	17	05	−06	10	20	75	08	08	−07	04	−12	09	09	01	70
66. Pertinent Questions (Part II) CMI	18	22	−09	07	20	68	17	−05	04	−01	11	07	17	12	69
67. Picture Gestalt NMT	14	13	06	25	12	23	24	37	18	03	−03	03	15	−04	42
68. Possible Jobs (Part I) DMI	12	05	11	07	08	20	82	16	01	07	12	08	02	19	84
69. Possible Jobs (Part II) DMI	24	14	04	06	03	21	56	15	09	06	27	08	09	18	60
70. Product Choice EMT	07	07	24	04	09	04	16	14	39	13	06	12	17	06	34
71. Seeing Problems EMI	.24	08	18	−11	08	43	11	22	27	05	09	−07	−15	−03	47
72. Sentence Selection EMI	27	07	16	20	−02	−02	15	09	13	05	25	10	03	54	56
73. Sentensense EMS	29	04	05	08	−01	22	11	−12	43	05	25	11	−02	29	52
74. Ship Destination Test CMS	08	13	02	50	05	01	−04	24	22	00	33	25	11	28	65
75. Similarities (Part I) CMT	11	05	07	10	53	25	07	00	16	−08	18	−04	17	16	50
76. Similarities (Part II) CMT	15	13	09	03	75	24	09	14	04	18	15	17	−01	05	18
77. Story Titles EMT	43	12	09	07	02	17	05	−05	09	17	36	25	01	17	50
78. Synonyms EMU	39	10	20	−09	03	14	00	11	12	11	12	05	20	35	45
79. Unlikely Things EMS	15	18	12	04	07	01	19	01	20	−01	03	55	27	25	59
80. Useful Changes EMT	28	−04	−06	−02	02	−04	−05	11	03	19	16	−03	55	07	47
81. Verbal Analogies I (Part I) CMR	26	18	06	26	14	−04	08	−08	07	22	50	10	10	14	55
82. Verbal Analogies I (Part II) CMR	24	22	43	09	07	−03	07	20	07	15	30	22	06	23	57
83. Verbal Analogies III EMR	19	19	06	15	04	−09	17	07	08	14	59	10	22	19	59
84. Verbal Classification (Part I) CMC	12	57	05	14	01	16	12	09	12	26	21	14	13	17	69
85. Verbal Classification (Part II) CMC	25	71	15	10	13	03	06	09	03	13	18	07	−01	28	76
86. Verbal Comprehension CMU	67	−04	06	04	10	14	01	12	09	08	21	08	26	21	68
87. Word Completion CMU	63	04	02	01	02	18	14	14	09	23	26	25	19	09	70
88. Word Extensions EMI	27	32	07	02	−06	04	11	11	06	04	37	16	22	39	58
89. Word Linkage EMR	20	13	60	13	10	−01	07	−06	19	18	00	09	18	11	57
90. Word Substitution EMU	60	15	−02	07	05	07	08	20	08	03	13	13	25	24	60
91. Word Systems EMS	09	01	15	24	07	10	−03	09	04	10	19	57	−11	10	51

Note.—Decimal points omitted.

the tests that are loaded on the CSU factor. It would appear from the test's history that the factor is concerned with the recognition of complete and correct words; a factor that might be called "word closure" (Pemberton, 1952). Such a conclusion is strengthened by the fact that Correct Spelling, designed as a measure of ESU, is also a measure of the recognition (evidently not in an evaluative way) of complete and correctly spelled words.

Tests for both CSU and NST were correlated in the previous analysis (Guilford, Merrifield, Christensen, & Frick, 1961) in which those factors were discovered. From Word Transformation's correlation with the CSU factor, although secondary, it appears that this symbolic redefinition task is dependent upon recognition of the symbolic units needed in effecting the transformation. The recognition of symbolic units necessary for sounding out words may account for the CSU loading for Sound Grouping.

CSC—Cognition of symbolic classes:

4.	Best Number Pairs (ESC)	.54
23.	Number-Group Naming (CSC)	.47
		(.38 NSS; .35 ESC)
21.	Number Classification (CSC)	.43
32.	Sign Changes II (ESR)	.35
		(.43 ESC)
3.	Best Number Class (ESC)	.32
		(.50 ESC; .32 ESS)
22.	Number Grouping (DSC)	.31
		(.34 EST)

The two shortened forms of the tests selected to measure the CSC factor function as anticipated, but the tests for CSC are led by a test designed for ESC. Like its analogous semantic test, Best Word Pairs, Best Number Pairs contributes more to the cognition-of-classes factor than to its parallel evaluation factor. Best Number Pairs asks *E* to evaluate which pair of stimuli makes the best class, but the specific properties of the best class, in other words, the specific criteria, are not defined in each item. Since the specific nature of the best class needs to be discovered for each item, cognition abilities should be expected to determine much of the test's variance.

The CSC factor defined in this analysis could be confined to the ability to recognize common properties of numbers, but a letter test has previously been loaded on it (Guilford, Merrifield, Christensen, & Frick, 1961). The significant loadings showing that three tests share relations to both CSC and ESC impressively demonstrate that it has not been easy to measure one of these factors without also measuring the other.

CSR—Cognition of symbolic relations:

29.	Seeing Trends II (CSR)	.46
		(.35 CMU)
16.	Jumbled Words (EST)	.35
		(.48 EST)
46.	Word Relations (CSR)	.31
		(.30 ESI)
43.	Word Choice (ESC)	.31
		(.32 CSI; .31 ESC)

Both CSR tests function as anticipated in this analysis. The significant CMU side loading for Seeing Trends II is not reasonable as the trends are based solely on the letter content of the words and not on their meanings. The result is consistent with a similar CMU side loading found by Guilford, Merrifield, Christensen, and Frick (1961), however. Some semantic recognition must somehow be involved in the test.

Depending upon anagrammatic rearrangements of letters of words, and the fact that some letters characteristically do or do not appear together, responding to Jumbled Words may involve CSR ability before evaluation enters the process. The common properties of many of the items of Word Choice were literal relationships—again indicating a need for CSR ability for responding.

CSS—Cognition of symbolic systems:

18.	Letter Triangle (CSS)	.39
33.	Similar Pairs (ESR)	.36
		(.35 ESR)
6.	Circle Reasoning (CSS)	.30

Letter Triangle leads the factor called CSS, the ability to comprehend systematic arrangements of symbols. The unusually small loading of Circle Reasoning does not strengthen this factor much. The implication derived from only slightly successful attempts to isolate CSS is that additional measures should be developed and analyzed so that stronger measures that consistently cohere are available.

Similar Pairs' loading on the CSS factor

should be accounted for by cognition that might be necessary to recognize the relations within word pairs. The determination of some of these relations is dependent upon alphabetical order, which also determines the systems in Letter Triangle.

CSI—Cognition of symbolic implications:

45.	Word Patterns (CSI)	.42
28.	S Test (ESI)	.41
35.	Symbol Grouping (CSI)	.33
43.	Word Choice (ESC)	.32
		(.31 CSR; .31 ESC)

The ability to foresee symbolic implications, CSI, is defined in this analysis by the two tests designed to measure the factor and also by the S Test, a test originally designed to measure sensitivity to problems. The original F test, from which the S Test was derived, probably did not indicate the traditional "sensitivity-to-problems" factor because that factor has been semantic. The semantic tests that did bring out the sensitivity-to-problems factor were later shown to belong largely on factor CMI, the cognition ability parallel to CSI, on which the S Test is now loaded. It appears that there is no difference between the ability to see symbolic implications and being sensitive to them.

CMU—Cognition of semantic units:

50.	SCAT Verbal (CMU)	.79
53.	PSAT Verbal (CMU)	.72
52.	ITED General Vocabulary (CMU)	.57
		(.31 ESS)
34.	Sound Grouping (ESC)	.38
		(.41 ESC; .30 CSU)
29.	Seeing Trends II (CSR)	.35
		(.46 CSR)
7.	Correct Letter Orders (ESS)	.32
		(.43 ESS)

The verbal-comprehension factor is clearly defined by the three tests selected to measure CMU. The tests are relatively univocal, having very high loadings on CMU.

Once again, Sound Grouping demonstrates its factorial complexity by splitting its variance between ESC, CSU, and CMU. Familiarity with the words used as stimuli for this test, in both their symbolic and semantic aspects, appears to facilitate either the pronunciation or the classification based upon the pronunciations. One might expect an analogous phenomenon underlying the CMU loadings of Seeing Trends II. But knowledge of meanings of the words in that test in no way aids in the discovery of the symbolic trend. The only common element in the CMU tests and Seeing Trends II is that words are employed.

MSI—Memory for symbolic implications:

24.	Numerical Operations (MSI)	.61
		(.33 NSI)
31.	Sign Changes (NSI)	.33
		(.30 NSI; .32 ESU)

Numerical Operations emerges clearly as test defining the numerical-facility factor. MSI receives further support from Sign Changes, a test designed to measure NSI, but which has consistently shared MSI variance. The shared variance on these two factors is not great enough, however, to cause serious concern regarding their distinctness.

DSC—Divergent production of symbolic classes:

40.	Varied Symbols (DSC)	.46

Varied Symbols once again serves as the principal measure of the DSC factor. Number Grouping, which was not loaded on DSC in a previous investigation (Hoepfner & Guilford, 1965), failed to be again. The strong face validity of Number Grouping as a measure of DSC apparently is misleading, as the test is complex in this analysis.

NSS—Convergent production of symbolic systems:

42.	Word Changes (NSS)	.56
		(.32 ESS)
23.	Number-Group Naming (DSC)	.38
		(.47 CSC; .35 ESC)
30.	Series Relations (ESS)	.31
		(.48 ESS)
25.	Operations Sequence (NSS)	.30

The two tests selected to measure NSS perform again in a reliable manner. The side loading of Word Changes on ESS might be due to a strategy that employs alternate orderings that are quickly evaluated according to the limitations imposed by criteria given in the test instructions. A similar kind of rationale, in reverse, could explain the minor secondary NSS loading of Series Relations. The *E* tries out each given operation in turn, producing a fully determined series.

NST—Convergent production of symbolic transformations:

5.	Camouflaged Words (NST)	.62
47.	Word Transformation (NST)	.43
		(.35 CSU)
3.	Best Number Class (ESC)	.31
		(.32 CSC; .50 ESC; .32 ESS)

The symbolic redefinition factor emerged in this analysis with both tests selected to measure it loading primarily upon it. The CSU side loading of Word Transformation was discussed in connection with the CSU factor.

NSI—Convergent production of symbolic implications:

14.	Form Reasoning (NSI)	.59
31.	Sign Changes (NSI)	.30
		(.33 MSI; .32 ESU)

The two tests selected to measure NSI perform as expected. To date, this solution represents the clearest separation between the MSI and NSI factors, although there still seems to be some common aspect in the tests used here to measure them and more univocal tests are apparently needed for both.

EFU—Evaluation of figural units:

26.	Perceptual Speed (EFU)	.69
15.	Identical Forms (EFU)	.63
11.	Derivations (ESU)	.31
		(.34 ESU)

The two EFU tests perform just as hypothesized, being univocal and highly saturated with common-factor variance. The EFU factor is the clearest interpretable factor to emerge in this analysis, probably due to its marked dissimilarity of content from the symbolic and semantic factors.

Interpretation of the Symbolic Evaluation Factors

ESU—Evaluation of symbolic units:

36.	Symbol Identities (ESU)	.62
19.	Letter "U" (ESU)	.56
11.	Derivations (ESU)	.34
		(.31 EFU)
20.	Marking Speed Test	.33
31.	Sign Changes (NSI)	.32
		(.33 MSI; .30 NSI)

The ESU factor emerged with remarkable clarity. The two speeded sensitivity tests that were previously suggested as measures of ESU (Guilford & Hoepfner, 1963) lead the factor with high loadings and little or no complexity. It is interesting to note that with strong tests of EFU and a sufficient number of tests for ESU, including Symbol Identities and Letter "U," in the analysis, the two factors separate very clearly. This decisive result clears up earlier uncertainties as to whether "perceptual-speed" tests composed of literal material should go with Thurstone's original perceptual factor or should represent a separate factor (Bechtoldt, 1947; Coombs, 1941; Thurstone, 1938b).

The third ESU test loaded on the ESU factor is Derivations, also a sensitivity test, involving words. Based upon the three ESU tests loading on this factor, it appears that ESU is the ability to make rapid decisions regarding the symbolic identity or accuracy of words, letter sets, and number sets. In Symbol Identities, there is a comparison of two given symbolic units to determine whether or not they are identical. In Letter "U," a word class is specified (words containing the letter "U"), with *E* to say whether or not each word satisfies the specification. Symbol Identities is a direct parallel to figural tests of EFU, in which figures are to be compared, with *E* to decide whether or not they contain exactly the same elements.

Derivations does not fit either of the two item models just described for Symbol

Identities and Letter "U." The things being compared are not exactly the same except for one element, as in the former, nor are class specifications given, as in the latter. The letters of the short word said to be extracted from the long word must coincide with a completely identical set of letters in the long word, except that the order is probably different. There is no clear model presented for comparison. This may be a reason for the lower ESU loading for Derivations than for Symbol Identities.

The two ESU "misses," those tests hypothesized for ESU but not loaded significantly on it (Correct Spelling and Familiar Letter Combinations), also aid in interpreting the ESU factor by indicating what ESU is not. One characteristic the two "misses" have in common but do not share with the other three ESU tests, is that the things with which comparisons must be made are not given on the printed page. They can only be compared with something in memory storage or something retrieved from memory storage, perhaps in the form of an image. In Correct Spelling, the needed model is the remembered correct spelling of each word. The task boils down to the question of how many of the 120 words in the test does *E* know, spelled correctly. This statement of the task makes it appear like a measure of cognition, as it turned out by analysis to be.

In Familiar Letter Combinations, the criterion for judgment is familiarity of the syllables or their observed probability of occurrence in *E*'s experience. In this test, there is no clear model for *E* to use, and what he has to use is also something in or from memory storage. Although it appears that the memory feature applies especially to the two ESU tests that missed, the question arises as to how general the implied evaluation principle is. If it is quite general, the principle that comparisons must be between perceived information would place an important restriction upon the definition of evaluation abilities.

The presence of Marking Speed, along with Sign Changes, here suggests some confounding of a finger-speed factor with both NSI and ESU. Rotation to an additional finger-speed factor might have cleared up the picture for both NSI and ESU. Implied is a general principle, namely, that the appearance of obliqueness among factors may be due to lack of a sufficient number of dimensions being included in orthogonal rotations.

ESC—Evaluation of symbolic classes:

3.	Best Number Class (ESC)	.50
	(.32 CSC; .32 ESS; .31 NST)	
32.	Sign Changes II (ESR)	.43
	(.35 CSC)	
34.	Sound Grouping (ESC)	.41
	(.38 CMU; .30 CSU)	
23.	Number-Group Naming (CSC)	.35
	(.47 CSC; .38 NSS)	
43.	Word Choice (ESC)	.31
	(.32 CSI; .31 CSR)	
37.	Symbol Manipulation (ESR)	.31
	(.59 ESR)	

Although three tests designed to be measures of the ESC factor are loaded on the factor called ESC, the factor is the least clear of all the new experimental factors found. The ESC tests are all complex.

The task involved in Best Number Class is to realize the numerical classifications of given numbers and then to select the one classification that is most valuable, value of number classes being defined by the test as the criterion.

Sign Changes II requires the examinee to change signs in a numerical expression so that the expression becomes an equation. Introspectively, it seems that a successful attack on such problems would include the tactic of becoming aware of what both sides of the expression have potentially in common, and from this awareness, to make the appropriate sign changes to bring about that common numerical value and thus change the expression into an equation. To clarify this attack with an example, consider the sample item: $3 + 1 = 6 \times 2$. The first step in effectively changing this expression into an equation is not to substitute signs, but to be sensitive to what the pair of numerical value 3 and 1, and the pair 6 and 2 potentially have in common. Their common, or class property is either the numerical value of 4 ($3 + 1$ and

6 − 2), or 3 (3 × 1 and 6 ÷ 2). Since only one of the necessary sign changes is given as an alternative answer, the only acceptable solution is the first one, and the common element in the expression is the value of 4, the value for which the signs must be changed. But this line of thinking suggests cognition rather than evaluation, factor CSC rather than ESC, and the loading for this test on CSC is only .35. Another hypothesis is that *E* somehow takes the offered solutions as *classes* of operation changes and considers them for adequacy.

The two ESC tests with the smaller significant loadings on this factor were constructed as estimation tests of ESC. They involve making a choice among given class properties on the basis of criteria supplied in the test. Although it would seem reasonable to have rotated the ESC axis to maximize its correlations with these more simple ESC tests, their tendency toward factor complexity and the loss of simple structure weighed against such a move.

ESR—Evaluation of symbolic relations:

37.	Symbol Manipulation (ESR)	.59
		(.31 ESC)
27.	Related Words I (ESR)	.43
33.	Similar Pairs (ESR)	.35
		(.36 CSS)

Three of the four tests designed to measure ESR, and no others, are loaded on the ESR factor, clearly defining it as representing the ability to make choices among symbolic relationships on the basis of similarity and consistency. The significant side loadings of the ESR tests were mentioned before in connection with the respective factors upon which the complex ESR tests are loaded.

ESS—Evaluation of symbolic systems:

41.	Way-Out Numbers (ESS)	.57
30.	Series Relations (ESS)	.48
		(.31 NSS)
7.	Correct Letter Orders (ESS)	.43
		(.32 CMU)
38.	Symbol Reasoning (ESI)	.41
3.	Best Number Class (ESC)	.32
		(.50 ESC; .32 CSC; .31 NST)
42.	Word Changes (NSS)	.32
		(.56 NSS)
8.	Correct Number Series (ESS)	.31
48.	ITED Verbal (CMU)	.31
		(.57 CMU)

Four of the five tests designed for ESS came out significantly loaded on ESS. The two leading tests, Way-Out Numbers and Series Relations, are in the estimation category and are composed of numbers. The two with distinctly smaller loadings, Correct Letter Orders and Correct Number Series, are in the sensitivity category, one being a number test and the other a letter test. It may be of some interest that while sensitivity tests proved to be better for ESU, estimation tests proved better for ESS. The trend must be better supported, however, before a principle can be stated. It is probably significant, however, that a system can deviate from a standard of comparison much more readily by degrees than can a unit.

Series Relations and Way-Out Numbers differ somewhat in the operations that they require. In the former, *E* probably makes new systems (series), following rules given in the alternative responses. He then compares each new system with the model that is given, deciding which new one comes nearest. The criterion is the degree of closeness of one set of numbers to another set. In the latter test, he is virtually to compare the distances of the first and last numbers in an irregular series to decide which one is farther from the other numbers. It is as if he were treating the same series first as one system and then as another, or from one point of view then from another. The criterion is numerical distance. The difference between these two relatively strong ESS tests contributes some breadth to the nature of the factor.

In terms of appearance, the two weaker ESS tests are close to being alternate forms of the same test, and both differ from the two stronger tests, as indicated above. They present letter series in the one case and number series in the other, with a verbal description of the principle that should be satisfied in a series. Sometimes

the series follows the principle exactly, sometimes not. It is probably significant that Correct Letter Orders has a significant loading on CMU whereas Correct Number Series does not. It matters more whether verbal terms are correctly understood in the letter test than in the number test. In the latter the rule can be more simply and precisely stated.

Symbol Reasoning would seem to be an ideal type of test for ESI, for it requires *E* to decide whether conclusions, expressed in symbolic form, can or cannot be justifiably drawn from other symbolic statements in the form of equations or inequalities. In essence, this test would seem parallel to the verbal-syllogistic test, Logical Reasoning, which had been found to measure factor EMI. It should be particularly interesting to learn why the symbolic form of the test was loaded instead on a systems factor.

The verbal test, Logical Reasoning, presents syllogisms, the relations or implications of which are contained in the premises and can be extracted by standardized tactics or rules of logic. The method of responding to the symbolic test, Symbol Reasoning, however, involves the estimation of an ordered series containing each element of the premises so that other relationships between values can be judged. It appears from this analysis that the construction (estimation) of a vaguely ordered number series is the process that discriminated most *E*s, and that is why the major variance of this test is shared with the ESS tests.

EST—Evaluation of symbolic transformation.

16.	Jumbled Words (EST)	.48 (.35 CSR)
10.	Decoding (EST)	.37
22.	Number Grouping (DSC)	.34 (.31 CSC)
39.	Typing Errors (EST)	.30
46.	Word Relations (CSR)	.30 (.31 CSR)

The three tests designed to measure EST emerged on the EST factor with unexpected clarity. The construction of the EST tests had proved to be most difficult and therefore it was expected that EST might not be found in this analysis.

All three tests involve the use of words, but they differ somewhat in terms of operations that *E* probably performs as he takes the tests. In Jumbled Words, he decides whether each response word could have come from the given word merely by rearrangement of the letters it contains. The criterion is in terms of a certain invariance or of element identity under the transformation of changed order.

In Decoding, *E* is expected to apply certain rules, of which five are given, in coding the letters of each of two words into a sequence of digits. The transformation in each case is in the form of substitution of elements according to rules. The *E* then is expected to decide which coding (transformation) result could be most easily and correctly decoded. The difference in ease and correctness of decoding depends upon the approach of the substitutions for a word to univocality, under the rules.

In Typing Errors, *E* is presented with what he is told is a misspelled word. From a knowledge of the arrangement of letters on the keyboard of a typewriter, which transformation in spelling (substitution of one letter element) has most probably occurred?

To summarize these comparisons, the transformation is in the form of reordering of letters in one test and in the form of letter substitutions and letter-digit substitutions in the other two tests. The criteria appear to be identity of elements in spite of transformation in one test, univocality of reference in the coding test, and nearness of position in symbolic system (keyboard) in the third. These differences suggest that there is some scope in kinds of transformations and in criteria for decision involved in connection with factor EST.

The two tests not hypothesized for EST, Number Grouping and Word Relations, share variance with EST tests possibly through common CST variance. CST has not been reported to be isolated as a factor, but the recognition that a change has occurred or the ability to see the changes

may be common to all five tests on this factor. The factor might therefore be a confounding of EST with CST.

ESI—Evaluation of symbolic implications:

2.	Best Letter Set (ESS)	.53
1.	Abbreviations (ESI)	.47
17.	Letter Problems (ESI)	.38

The ESI factor is defined by two tests designed as measures of ESI, but is led by a test designed for ESS that is not loaded on that factor. The most obvious common characteristic of the two leading tests is that the given stimuli are sets of letters, but this is also true of many other tests in the battery. But in addition, the item formats of the two tests are similar. In Abbreviations, a sequence of three or four letters is presented as the potential abbreviation of three alternative, familiar words, with *E* to say for which word the abbreviation best stands or that it best implies. In Best Letter Set, a sequence of three of four letters is given and three alternative letter sets of the same length, *E* to select the one that is most similar to the given set by virtue of common properties. The test was expected to be a measure of ESS because it was thought that the nearness of one set to another in terms of their constitutions would be a matter of comparing systems for approach to identity in terms of systemic properties.

It is not very clear how Best Letter Set becomes an implications test rather than a systems test. It is little more than stating the obvious to say that the given set apparently implies the best alternative, similarly to the way in which a letter set in Abbreviations implies a longer letter set in the form of a real word. A revised format of Best Letter Set, giving alternatives that either do or do not fit the principle of the given set exactly, might have been a better ESS test. But there would still be much unanswered about underlying reasons for the difference in factor content of the two test formats.

Letter Problems holds some promise of univocality, so far as this analysis goes, but with a low loading of .38 and a reliability of .88, there is considerable room for additional common-factor content. Letter Problems is interesting for the fact that is was developed by analogy to Form Reasoning, which is a marker test for factor NSI. A major difference, which should not be factorially significant, is that Letter Problems uses letters as symbolic elements whereas Form Reasoning uses geometric forms as symbolic elements. The significant difference is that Form Reasoning calls for inferences or conclusions to equations of a certain type whereas Letter Problems asks for decisions as to whether a problem equation is solvable, or is solvable with a minor change in the problem. It might be said that the criterion for evaluation is solvability or the possibility of valid inferences or implications. This is not strictly a matter of judging the value or identity of an implication, as such, or its similarity to another implication, which are common kinds of criteria in tests of other evaluation factors. But if a new kind of criterion is involved and is crucial to the loading on ESI, we have a little extension of connotation of evaluation abilities and the evaluative process as envisaged from the psychometric approach.

Interpretation of the Semantic Reference Factors

CMU—Cognition of semantic units:

51.	CAT Reading Vocabulary (CMU)	.70 (.47 CMS)
86.	Verbal Comprehension (CMU)	.67
87.	Word Completion (CMU)	.63
90.	Word Substitution (EMU)	.60
77.	Story Titles (EMT)	.43 (.36 EMR)
52.	CAT Arithmetic Reasoning (CMS)	.40 (.57 CMS)
78.	Synonyms (EMU)	.39 (.35 EMI)
64.	Matched Verbal Relations (EMR)	.32 (.50 EMR; .34 EMI)
57.	Class Name Selection (EMC)	.31 (.50 EMC)

In factor analyses of semantic-test batteries, it is not surprising to find a number of tests loaded on the CMU factor. Word knowledge is the essence of factor CMU—the well-known factor of verbal comprehension. The achievement tests appear to be factorially complex, spreading their variances over CMU and CMS, the factors most important for academic performance.

The Verbal Comprehension test used in this study was so constructed as to eliminate some of the items of standard verbal comprehension tests that would appear possibly to introduce evaluative variance into the scores. In Word Completion, *E* is to write synonyms or short definitions for given words. The test was developed to test the hypothesis that a completion form of vocabulary test would yield a more pure measure of CMU and that a multiple-choice form may have some secondary evaluation (EMU) variance. The correlation between Verbal Comprehension and Word Completion is .68, and both tests are found to be strong univocal measures of CMU, with the multiple-choice form having more CMU variance.

Word Substitution and Synonyms were both designed for the EMU factor. In Word Substitution, *E* is to select a word that best fits into a given sentence as a substitute for an underlined word. The alternative choices are all about equally acceptable, so that *E* is required to exercise evaluative thinking to find the best answer. Synonyms is also different from traditional vocabulary tests in that the choices are about equally acceptable and the difficulty of the words in each item is at a low level in order to reduce CMU variance in the test scores. In spite of the attempt to minimize cognitive variance and to maximize evaluative variance, both tests have significant loadings only on CMU, indicating that the evaluative process cannot be emphasized over the cognitive process in the case of semantic units merely by increasing the competition among alternative choices.

The outcome with respect to Word Substitution and Synonyms answers an important question regarding what kind of test measures factor CMU. Most vocabulary tests demand only that the examinees show acquaintance with words or some familiarity with them. There is usually no demand for very penetrating or discriminating knowledge of words. The presence of the two EMU-designed tests in the CMU lists indicates that ability to make fine discriminations among more familiar words is also a matter of factor CMU.

The finding that CMU is also indicated by tests calling for fine discriminations between word meanings could also account for the appearance of some CMU variance in other verbal tests where the vocabulary level has been kept low with the objective of reducing the CMU variance. Although the vocabulary level is low, some moderate or high-level discriminations on the basis of connotative aspects of word meanings may be involved, such as in Story Titles, Matched Verbal Relations, and Class Name Selection.

CMC—Cognition of semantic classes:

85. Verbal Classification—Part II (CMC)	.71
84. Verbal Classification—Part I (CMC)	.57
56. Best Word Pairs (EMC)	.34
88. Word Extension (EMI)	.32
	(.39 EMI; .37 EMR)

The CMC factor appears to be well defined in this study. Since variables 85 and 84 are two parts of the same reference test, the factor loadings of these variables are likely to be overestimated, being somewhat inflated by variance specific to the test Verbal Classification.

As in the symbolic study, the evaluative classes test that did not specify the criteria for evaluating class membership is loaded on the cognition factor. This behavior of analogous tests shows strikingly the necessity for specified criteria for evaluative processes to occur.

CMR—Cognition of semantic relations:

89. Word Linkage (EMR)	.60
82. Verbal Analogies I—Part II (CMR)	.43
	(.30 EMR)

Although Word Linkage was developed to measure the EMR factor, it was found to be the leading variable on what appears to be CMR. In each item of Word Linkage, E is to choose a word that is related to two other given words in different ways. The present finding seems to indicate that the task is heavily dependent upon cognitive abilities because the important difficulty lies in *seeing* the two different relationships, and this depends also upon E's having rich meanings for the words involved, hence the CMR variance.

Verbal Analogies I, as administered in this study, has three equal, separately timed parts. For the process of factor analysis, two part scores were wanted, so variable 82 is a combination of the second and the third parts, and variable 81 is the first part. Variable 82 contributes consistently to the CMR factor as expected, but variable 81 is loaded significantly on the EMR factor. The reason for this difference in behavior of the parts of Verbal Analogies I is not obvious. The CMR factor is not as sharply confirmed as was expected. The tests that are loaded on this factor, however, suggest that the factor is similar to the CMR factor isolated in the past.

CMS—Cognition of semantic systems:

52.	CAT Arithmetic Reasoning (CMS)	.57 (.40 CMU)
74.	Ship Destination Test (CMS)	.50 (.33 EMR)
51.	CAT Reading Vocabulary (CMU)	.47 (.70 CMU)

This is the well-known factor of general reasoning (Kettner et al., 1956). Ship Destination Test and Arithmetic Reasoning have been the two leading tests defining this factor. The heavy involvement of one measure of academic achievement in the leading test probably accounts for the "pulling" of variance from Reading Vocabulary into CMS. When complex standardized tests have been employed in factor analyses, they have tended to be complex.

CMT—Cognition of semantic transformations:

76.	Similarities—Part II (CMT)	.75
75.	Similarities—Part I (CMT)	.53

The two parts of Similarities clearly define this reference factor. Owing to the fact that these two variables are two parts of the same test, their loadings are probably overestimated.

CMI—Cognition of semantic implications:

65.	Pertinent Questions—Part I (CMI)	.75
66.	Pertinent Questions—Part II (CMI)	.68
71.	Seeing Problems (EMI)	.43
53.	Apparatus Test (EMI)	.34

The CMI factor was originally called "conceptual foresight" (Berger et al., 1957). Two parts of Pertinent Questions are the leading variables defining this factor in the list above. Again, there is inflation of loadings from the specific-variance source.

In previous studies, Seeing Problems and Apparatus Test, one or both, commonly helped to define a factor called "sensitivity to problems," which was defined as the "ability to see defects, needs, and deficiencies," and was considered to belong to the category of evaluation. In this study these two tests were loaded significantly together on factor CMI. Seeing Problems and Apparatus Test have frequently had variances from cognitive factors, sometimes from factor CMI and sometimes CMT. Whether the other tests that have helped to determine the sensitivity-to-problems factor in the past—Social Institutions and Seeing Deficiencies—will also be found consistently related to either CMI or CMT, or both, is yet to be determined.

DMI—Divergent production of semantic implications:

68.	Possible Jobs—Part I (DMI)	.82
69.	Possible Jobs—Part II (DMI)	.56

The DMI factor has been isolated a number of times in previous studies. Since variables 68 and 69 are two parts of the same test, their factor loadings are proba-

bly overestimated on this factor. No evaluation tests appear to be related to it. Whether evaluation factors in general are as easily differentiated from parallel divergent-production factors remains to be seen. There had not been much concern about such discriminations, as indicated by the scarcity of other divergent-production reference factors in the study. The exception to this lack of concern was some possible confusion between DMI and EMI (sensitivity to problems) as seen in one study (Guilford, Merrifield, & Cox, 1961), where some tests shared variances from the two factors. In general either divergent- or convergent-production tests require the *producing* of responses to fit given data in various ways, whereas in evaluation tests responses are *presented* to *E* for evaluation purposes.

NMT—Convergent production of semantic transformations:

67.	Picture Gestalt (NMT)	.37

Picture Gestalt, a test designed for the NMT factor, has identified a factor, presumably NMT, in agreement with a previous study (Wilson et al., 1954). Picture Gestalt had been put into the battery to help bring out NMT and to see whether judgment tests would show any relations with NMT, as one judgment test had in the past. They did not show that relationship in this study.

Interpretation of the Semantic Evaluation Factors

EMU—Evaluation of semantic units:

61.	Double Descriptions (EMU)	.64
73.	Sentensense (EMS)	.43
70.	Product Choice (EMT)	.39

The EMU factor was hypothesized as an ability to evaluate the suitability or adequacy of a word or object in terms of meeting given criteria. In each item of Double Descriptions, *E* is to evaluate the extent to which objects possess two criterion properties. The nature of the task in this test seems quite congruent with the conception of the definition of EMU.

Two other tests designed for factor EMU (Word Substitution and Synonyms) are missing from the list of tests above. Instead, they were loaded substantially on CMU, the *cognition* of semantic units. It is not surprising that essentially modified vocabulary tests should have some loading on CMU, but it is surprising that with the words of a fairly high level of familiarity and with the alternative answers designed to emphasize uncertainty of choice, there was not at least significant evaluative variance.

Sentensense was originally designed for factor EMS, since it deals with the internal consistency of ideas or events expressed in a sentence, which is defensible as one of the forms of semantic system. The test was developed by analogy to Unusual Details, which led in defining what was believed to be factor EMS previously. In both, inconsistences are to be noted. The significant loading of Sentensense on EMU may mean one of two things. One hypothesis is that the soundness of ideas or events expressed in a sentence was evaluated taking each sentence as an integrated whole or unit rather than as a system of interrelated parts. In other words, the sentences were possibly treated as semantic units in the psychological processing of information. The shortness of the sentences in Sentensense and the relatively high ability level of the *E*s in this study might have made this possible. But each sentence contains two ideas or events, which suggests another hypothesis. The unit may have been each component of the sentence rather than the entire sentence. Decisions were made regarding the compatibility of the pairs of ideas or events as pairs of units. In the case of either hypothesis, an inference might be that more complex sentences might have shifted the test where it was expected, to the EMS factor.

Product Choice was designed for factor EMT. In each item, from a set of alternatives *E* is to select an object or objects that can be used most adequately for a specified purpose. In order to find the right answers, *E* must evaluate how adequately each object is used in an unconventional

way. The significant loadings of this test on EMU, however, indicate that the test has little to do with the psychological phenomenon of transformations. After all, it is the *objects* that are to be evaluated, not the transformations as such. Such items of information are units.

EMC—Evaluation of semantic classes:

55.	Best Word Class (EMC)	.54
57.	Class Name Selection (EMC)	.50 (.31 CMU)
58.	Commonsense Judgment I (EMT)	.33 (.33 EMI)
62.	Important Facts (EMS)	.33

Two of the three tests developed for EMC, Best Word Class and Class Name Selection, are the two leading tests defining the factor. Best Word Class deals with the evaluation of class names offered to represent given *single* objects. Class Name Selection deals with the evaluation of names given to represent *sets* of objects.

Best Word Pairs, a test developed according to the EMC hypothesis, contributed instead to factor CMC. The *E* is asked to evaluate pairs of words in each item, saying which pair makes the best class. Since the nature of the class is not given, *E* has to discover the nature of each class for himself, which could account for the cognition variance in factor CMC. Also probably unfavorable for determining evaluation variance in the test is the fact that the criterion "best" was not defined sufficiently precisely.

Commonsense Judgment I was predicted for the EMI factor, but it shares its variance equally with EMC. Commonsense Judgment I has to do with the evaluation of possible faults in a proposed plan. It is difficult to see how genuine classes are involved in this test.

Important Facts was designed for the EMS factor based upon the hypothesis that a problem situation is a kind of semantic system. The hypothesis was derived from several previously analyzed tests for the CMS factor, including Ship Destination Test and Arithmetic Reasoning, both of which require *E* to structure problems. Important Facts requires *E* to judge the relative importance of a set of given facts in connection with solving simple problems. However, Important Facts was found loaded on EMC and not on EMS, suggesting that what is evaluated in the test is something with class properties. The involvement of classes in this test is not obvious.

EMR—Evaluation of semantic relations.

83.	Verbal Analogies III (EMR)	.59
64.	Matched Verbal Relations (EMR)	.50 (.34 EMI; .32 CMU)
81.	Verbal Analogies I—Part I (CMR)	.50
54.	Best Trend Name (EMR)	.48
88.	Word Extension (EMI)	.37 (.39 EMI; .32 CMC)
77.	Story Titles (EMT)	.36 (.43 CMU)
74.	Ship Destination Test (EMS)	.33 (.50 CMS)
63.	Logical Reasoning (EMR)	.32 (.40 EMI)
82.	Verbal Analogies I—Part II (CMR)	.30 (.43 CMR)

From the structure-of-intellect model, the EMR factor was hypothesized as the ability to evaluate relations between words or ideas. In accordance with the hypothesis, four tests were newly developed or adapted from existing tests. All four are represented in the list above. Verbal Analogies III and Matched Verbal Relations were adapted from the usual tests of verbal analogies, which emphasize the cognition of relationships between given pairs of words. In tests of EMR, however, the relations in the first pairs of words are made obvious, in order to minimize the variance due to *E*'s ability to cognize the relationships. On the other hand, the choices of alternative completions are made difficult, in order to maximize the evaluative variance by requiring *E* to compare the given alternatives in the light of the standard relations specified in the first pairs. The fact that Verbal Analogies III has a significant loading on EMR and has

no side loading illustrates the possible separation of evaluative ability from cognitive ability by imposing a difficult operation at the appropriate phase of the verbal-analogies task.

The well-known Logical Reasoning test was included in this study to provide continuity with the factor of logical evaluation identified in the previous studies. The present findings suggest that the logical-evaluation factor is not the same construct as the EMR factor in this study.

Part I of Verbal Analogies I, a marker test for the CMR factor, has significant loadings on the EMR factor. It appears that Part I of this test is not a measure of the CMR factor, but has significant evaluative variance. Part II of Verbal Analogies I restricts itself more to factor CMR, however.

The small but significant EMR loadings of Word Extension and Story Titles probably result from the difficult relationships between the given items and the alternatives. Ship Destination Test has a long history of being a pure measure of general reasoning, and no explanation of its EMR loading is apparent.

EMS—Evaluation of semantic systems:

91.	Word Systems (EMS)	.57
80.	Unlikely Things (EMS)	.55

Word Systems and Unlikely Things are univocal tests for this factor. In Word Systems an item is composed of three matrices with approximate meaningful sequences of words in the columns and rows of each matrix to be judged for the best ordered system. Unlikely Things was adapted from a test called Unusual Details which asks *E* to find unusual items of information in sketches of common situations. The EMS factor identified in this study is similar to the factor called "experiential evaluation" defined by Unusual Details by Hertzka et al. (1954), but the conception of its nature is broadened considerably by the relation to the test, Word Systems.

Two other tests designed for EMS, Important Facts and Sentensense, have been mentioned above. In retrospect, the present knowledge concerning the CMS and NMS factors have led to a number of hypotheses concerning the nature of the EMS factor. The present findings lead to reservations concerning the possible approaches based upon (*a*) evaluation of internal consistency of a two-idea sentence, and (*b*) evaluation of importance of facts needed to solve a given problem. It may be that in a test for EMS, the entire conceived problem should be the object of evaluation, not a particular detail of the problem, or that internal consistency of the stated facts should be emphasized in such a test.

EMT—Evaluation of semantic transformations:

80.	Useful Changes (EMT)	.55

Although only a singlet, Useful Changes was rotated so that its unique variance was orthogonal to the rest of the factor matrix. This is the weakest evaluation factor in this study, but it was felt that the weak factor could offer guidelines about EMT test construction and also shed some light on the nature of the factor.

Product Choice, a test designed for EMT, was loaded instead on EMU. In the interpretation of EMU, it was suggested that the objects presented as alternative choices in this test were the products offered for evaluation rather than the transformations, as such.

In the area of divergent production, the tests of transformations require *E* to produce numerous transformed ideas. On the other hand, the tests of transformations in convergent production require *E* to produce single transformed ideas in order to attain specified goals. Two alternative approaches in developing the tests for evaluation of transformations were considered. The first approach is represented by Product Choice and Useful Changes, in which the task is to evaluate the objects to be transformed to attain a specified goal. The second approach is to require *E* to evaluate the results of transformations or the transformed information in the light of uniqueness or adequacy as the criterion.

Story Titles, another test designed for

EMT, is based upon the second approach. The test requires *E* to choose the best titles that give new interpretations for short stories. The analyses indicate that Story Titles failed to identify the EMT factor; instead, the test shared substantial variance with CMU.

To understand why Useful Changes, a first-approach EMT test, is loaded on EMT while Product Choice is not, one must look closely at how *E* processes the test information in responding. In Product Choice *E* is given two objects and is asked to decide what product from among the alternatives could be best made. Each product-alternative is then evaluated according to the limitations inherent in the two given objects—a process much like that of Double Descriptions.

Useful Changes is a reversal; *E* is given a task to perform and is to decide which of three given objects could best be transformed to perform the task. In other words, *E* must transform (or attempt to transform) each object and then judge which transformed object would perform the job most adequately. It is the transformation that is evaluated, not the object. Future tests for EMT should demand *E*'s actually making or recognizing simple transformations, and then judging them according to some criteria.

EMI—Evaluation of semantic implications:

60.	Complete Thoughts (EMS)	.55
72.	Sentence Selection (EMI)	.54
63.	Logical Reasoning (EMR)	.40
		(.32 EMR)
88.	Word Extensions (EMI)	.39
		(.37 EMR; .32 CMC)
78.	Synonyms (EMU)	.35
		(.39 CMU)
64.	Matched Verbal Relations (EMR)	.34
		(.50 EMR; .39 CMU)
58.	Commonsense Judgment I (EMI)	.33
		(.33 EMC)

Complete Thoughts, a test designed for factor EMS, is the leading test defining EMI. The test calls for decisions as to whether statements are complete or incomplete sentences, that is, whether implications generated by the beginnings of sentences are fulfilled in the remaining words. It would appear that the alternative viewpoint of the implication-fulfillment aspect of sentences, completeness, does not apply to the evaluation of systems, for which Complete Thoughts was designed.

The second strong EMI test, Sentence Selection, requires *E* to select the statement that is most probably true, in view of the given information, a single premise. The test has high face validity in reference to the adopted conception of EMI, which is hypothesized as an ability to evaluate extrapolated information in the form of expectancies, predictions, antecedents, concomitants, or consequences.

The well-known test Logical Reasoning, a syllogistic test, has its highest loading on EMI, but shows some sign of factorial complexity, with a significant loading on EMR. Apparently, the relational aspects of premises and conclusions are not strong enough to emphasize the product of relations. Conclusions are more like expectancies, following from the premise.

Word Extensions, a test designed for factor EMI, has significant loadings on the EMI and EMR factors also. The test asks *E* to select the name of an object or attribute that is always implied by a given thing. Some of the easy items in this test might have been answered by evaluating the closeness of relationship between objects and the given thing, hence the relation of the test to EMR. More obvious or more meaningful implications may be psychologically regarded as relations. Without further evidence concerning each item of this test and its relation to EMI or EMR, it is best concluded that *Word Extensions* has loadings on both the EMI and EMR factors.

The fact that factors EMR and EMI share three tests in common in this analysis might suggest some obliqueness for these two factors. But it should be noted that each of these factors has at least three other tests not shared significantly by the other. The hypothesis of orthogonality cannot therefore be given up. Factors EMI

and CMU also have two tests in common, but more than two for each factor that are not shared significantly.

Discussion

In the two studies reported, 12 factors of evaluative abilities were deduced from the structure-of-intellect model, and the validity of the hypotheses was tested by factor analysis. The hypothesized factors pertain to the evaluative operation and require tests having symbolic and semantic content. The hypotheses for the two studies, couched in a factor-analytic design, would read as follows:

1. The six symbolic-evaluation abilities predicted by the structure-of-intellect model are essentially independent of one another and also of other known factors of intelligence, particularly those of symbolic cognition.

2. The six semantic-evaluation abilities predicted by the structure-of-intellect model are essentially independent of one another and also of other known factors of intelligence, particularly those of semantic cognition.

Mutual Independence of the Evaluation Factors

Of the 12 hypothesized evaluation factors, the independent existences of 11 have been well supported by the fact that the leading variables defining them have few significant side loadings on other factors of evaluation; what secondary loadings there are, are usually small and they are on cognition factors.

Factor ESU, defined best by Symbol Identities and Letter "U" is clearly the symbol evaluation factor which had never emerged clearly before, due to inadequate numbers of tests for its delineation from perceptual speed. ESU tests have some small perceptual-speed involvement, but this could be due to the speeded natures of the tests defining the two factors. ESU is defined as representing the ability to judge quickly and accurately literal and numerical information as conforming or not conforming to certain necessary criteria and is probably equivalent to the ability commonly known as clerical speed and accuracy.

Led by Best Number Class and Sign Changes II, both somewhat complex, ESC appears to represent the ability to judge the goodness of class membership of symbolic information and the ability to be sensitive to and judge the class properties inherent in symbolic expressions. The essential nature of this factor lies in evaluating the "tightness" of the concept embracing the given symbolic information. Such an evaluation is necessary for judging concept names and for judging class membership. This ability has been found to be involved to a significant degree in success in high-school mathematics (Guilford et al., 1965).

The ESR tests defined a factor involving the judgment of relationships among symbols on the basis of similarity and consistency. The leading test, Symbol Manipulation, calls for judgments regarding the truth or falsity of immediate consequences of symbolic propositions. The immediacy of the consequences is the important aspect for the relational nature of this test.

Somewhat broader than a numerical estimation factor, ESS reflects the ability to estimate values of symbols within a vaguely ordered system. The leading tests, Way-Out Numbers and Series Relations, support this interpretation for numerical series, but other tests define the factor more broadly to include sensitivity to errors within the symbolic systems, thus, that nonoperational ability, sometimes referred to as "number sense" or "approximation ability," finds a place in the structure of intellect as a unique, measurable factor of symbolic evaluation.

The EST factor isolated was also broad in meaning, encompassing sensitivity to symbolic substitutions and reorderings. It is probable that tests of a more cryptographic nature, perhaps based upon universal characteristics of codes (letter placements and frequencies) will function as stronger and purer measures of EST. If this is the case, another high-level job element will have been accounted for in the theoretical model of human intelligence.

The clear factorial delineation of ESI is aided (and complicated) by what is probably common format variance. The connections to be evaluated in Best Letter Set, Abbreviations, and Letter Problems are more remote than those evaluated in the relations tests. Whether the remote nature of the connections to be evaluated is necessary for the product of implications must await further investigation. It is an interesting conjecture that evaluation of implications is a rather probabilistic decision-making process.

EMU is clearly defined as the ability involved in judging the suitability or adequacy of ideas and objects in terms of meeting certain criteria. In all three tests loading on EMU, Double Descriptions, Sentensense, and Product Choice, the criteria are given and the ideas or objects are evaluated in terms of how they meet those criteria.

Best Word Class and Class Name Selection identify factor EMC as being almost exactly parallel to the symbolic factor, ESC. The criteria for judgment is the "tightness" of the concept or class and the things to be judged are either members of the class or concept names.

The EMR factor, as represented by Verbal Analogies II and Matched Verbal Relations, is also closely parallel to the ESR factor and is concerned with judgments regarding the similarity or consistency of relationships between words or ideas. Like ESR, the EMR factor also embraces relationships of a consequences type, when the connections are very immediate. In distinguishing relations and implications, we may consider the product of implications as a broad category of connections between ideas, some of which, when specifiable because they have unique characteristics, are processed as relations.

The factor of logical evaluation has been defined as an ability involving sensitivity to consistency of logical relationships and has been consistently defined by tests of decisions about the correctness of conclusions drawn from premises. The factor was identified with EMR in the SI model, with the thought that each statement, premise, or conclusion expresses a relationship of some kind. But from another point of view, conclusions are inferences, and inferences are implications, and one could say that in tests such as Logical Reasoning, implications are being judged. This line of thinking led to doubts about identifying logical evaluation with EMR rather than EMI. The results are rather decisive that other kinds of tests, especially tailored for EMR, determine a factor that has a better claim to that spot in the SI model.

The EMS factor as defined by Word Systems and Unlikely Things is probably broader than what can be confidently concluded from the nature of its two tests. In any case, it is much broader than the old factor of "experiential evaluation," but probably subsumes the old concept, in part. The essential nature of this factor is concerned with judgments about or within complexes of meaningful information, where the complex is a necessary consideration for judgment. Where the stimuli are not complex, but can be processed as simple information, as in Sentensense, the evaluation is of a unit, not a system.

This study resulted in revealing only a trace of the factor EMT, defined by Useful Changes. Future tests for this factor must employ transformation as the information to be evaluated, not transformed objects. This principle is difficult to follow using paper-and-pencil tests, but Useful Changes should serve as a prototype of the tests that might fruitfully be tried.

The EMI factor may be regarded as a new factor, represented best by Complete Thoughts and Sentence Selection. It is definitely not the former sensitivity-to-problems factor, whose tests now appear to be cognitive and to belong with factor CMI. Nor is it clearly the former "logical evaluation," defined mainly by the test Logical Reasoning, and other tests of similar nature, although there is something in common to the two factors, with Logical Reasoning furnishing a link. It should be noted that the factorial behavior of Logical Reasoning is very similar to that of its parallel symbolic test, Symbol Reasoning. The probabilistic nature of the ESI factor,

however, does not appear well founded as an interpretation for EMI, where the implications can only be defined as being relatively remote.

The cell for EMI in the SI model was previously assigned to the factor called *sensitivity to problems*. Two reasons arising from this study call for rescinding that decision. One, just mentioned, is the finding of a more suitable factor for that spot. The other is that the tests used as markers for the sensitivity-to-problems factor accommodatingly went together elsewhere, joining with the marker tests for CMI, the *cognition* of semantic implications. The change from EMI to CMI is a change in operation only. Sensing problems and seeing implications are compatible ideas, at least.

Independence of the Evaluation Factors from the Reference Factors

All of the five kinds of operations predicted by the structure-of-intellect model were involved in these studies, but obviously not equally so. Outside the area of evaluation under primary consideration, there were 11 different cognition factors represented by marker tests, 1 memory factor, 2 divergent-production factors, 4 convergent-production factors, and 1 figural-evaluation factor. The reference factors were included in the analysis with the usual concern lest some of the new evaluation factors not be shown to be distinct from parallel factors in other operation categories or lest some of the variances of new experimental tests not be well accounted for.

The greatest concern was regarding the demonstration that the evaluation abilities be differentiated from the parallel cognition factors. In constructing evaluation tests, it is not easy to be sure that cognitive variance has been ruled out by the experimental controls in the test conditions, or even that cognitive variance may not dominate the test. Other reference factors were brought in because new experimental tests were suspected of involving operation variances other than evaluation.

Evaluation was defined earlier as the process of reaching decisions or making judgments concerning the goodness of information in terms of specified criteria. It was thought that one unique characteristic of good evaluation tests, as distinguished from cognition tests, would be the condition of uncertainty at the point of the response to alternative choices rather than at the point of knowing the specifications of the criteria. In developing the tests of evaluation, attempts were made (*a*) to specify the criterion for evaluation as clearly as possible, (*b*) to induce the uncertain condition by making alternative choices about equally "good" (the "uncertainty principle"), and (*c*) to keep the difficulty of words in the items at low levels. Although these principles appeared to work in some instances, in the measurement of evaluation, there are several instances in which tests designed for evaluative factors, following the principle of maximizing uncertainty, are loaded on factors of cognition. Careful observation of these tests might throw light upon the nature of the evaluative processes. It might also serve to point out some of the problems in test development.

Word Substitution and Synonyms were both designed for the EMU factor, emphasizing the uncertainty principle. In Word Substitution, *E* is asked to select a word that best fits into a given sentence in place of an underlined word. Synonyms is similar to the usual multiple-choice vocabulary tests except that the alternative choices are all about equally acceptable and that the difficulty of words, as such, is kept at a low level. In spite of the attempt to maximize evaluative variance and minimize cognitive variance in this manner, the analysis indicates that both tests are essentially CMU tests. The competing choices in Synonyms and Word Substitution probably increased the CMU variance by requiring *E* to exercise finer discrimination among the meanings of words. The fact of competing choices alone does not ensure evaluative variance.

One of the important differences between these two tests and Double Descriptions, the leading test for EMU, is that in the latter the criteria for evaluation are explicitly stated in the items whereas the criteria for Word Substitution and Synonyms

are roughly stated as the adequacy of a word in a given sentence, or closeness of the meanings of words.

In both Word Substitution and Synonyms, the criteria are not explicitly stated in connection with each evaluative act. In either case, *E* has to cognize his own criterion, which is the exact meaning of the given word. Success in this step is crucial for success in answering the item. A reasonable conclusion is that once *E* has defined the words precisely, there is not much uncertainty as to which match is best.

A similar factor switch occurred on the parallel symbolic factor, ESU. The test Correct Spelling does not provide specific criteria for reaching decisions other than drawing upon memory storage for something with which to compare the given item of information. Correct Spelling is loaded on CSU, the parallel cognition factor.

In other forms of tests, the specifications of criteria do not seem to present a serious problem. In each item of Verbal Analogies III, for example, the standard pair of words is given as a criterion followed by the first word of the second pair and the alternative choices. First, *E* is to cognize the relationship in the standard pair of words, a step that is made fairly easy for him. The second step is fitting the first word of the second pair into the criterion relationship discovered in the first pair. This step can be identified as the construction of a search model. The third step is to try out the alternative words to see which one best fits the search model. In construction of this test, there was an attempt to introduce the uncertain state of affairs in the third step in order to increase evaluative variance. In other words, it is where the uncertainty is encountered that matters for testing purposes, not the fact that there is uncertainty.

In general, the strategy in test development employed in these studies was to devise tasks in which only the evaluative process is difficult in the total problem-solving activity. The results seem to suggest that certain approaches found to be promising in one test are not necessarily applicable in other tests, and certain approaches found to be promising in one product category of the structure-of-intellect model do not necessarily apply in other product categories. There is sufficient evidence, however, that evaluative abilities can be measured separately from cognitive abilities by presenting adequate specifications of criteria for evaluation and providing the appropriate level of uncertainty in connection with the response to the alternative choices.

A second cognition-evaluation problem was cleared up in both studies. As the semantic sensitivity-to-problems tests, hypothesized for EMI, were loaded instead on CMI, so did the parallel symbolic test. It appears that sensitivity to problems is similar to foresight—one who plans well is one who is sensitive to potential problems.

A third example occurred with the classes tests of both studies. Tests of ESC were direct translations of the EMC tests into symbolic content. Best Number Pairs was an adaptation of Best Word Pairs. The two performed very similarly in that both went rather on their respective cognition parallels, factors CSC and CMC. In both cases it can be pointed out that for measurement of evaluation rather than cognition, there is a need to state explicitly the criteria upon which judgments are to be made, and possibly provide models for comparison or to describe them, as in Double Descriptions and Letter "U."

A more general indication of the overall cognition-evaluation confusion is indicated by the frequency with which tests designed for one operation have significant loadings on factors in the other operation category. In both studies combined, the hypothesized cognition tests exhibited 28 loadings on cognition factors and 6 loadings on evaluation factors, implying that evaluation does not account for much of the variance in cognition tests. Unfortunately, the converse is not so strikingly clear. The evaluation tests had 22 loadings on cognition factors and 48 on evaluation factors, indicating that cognition does account for some of the variance in evaluation tests.

It would appear to be easier to keep evaluation out of cognition tests than to keep cognition out of evaluation tests, although

this depends somewhat upon where the rotations happen to go in a particular analysis. At any rate, from another point of view the experimental evaluation factors appear to be well differentiated from other operation factors. Although one might expect evaluation to play roles in connection with convergent-production tests, in which responses must be narrowed down to a single one for each item, as far as can be seen there is little reason to doubt the distinctness of those two operation categories. The tests of memory, divergent production, and convergent production exhibited their loadings mainly on their respective factors, and the factors were relatively clear of evaluation-test loadings. The expectation that such confusions would not be great was confirmed.

The Properties of Evaluation

What have the studies done for further elucidation of the concept of the operation of evaluation? What aspects of the concept possibly need changing as a result of the new information about factors and their tests, and what new features come into the picture? The answers to these general questions shall be considered more specifically in terms of (*a*) the kinds of judgments that belong in the picture of evaluation as defined by the tests; (*b*) what kinds of criteria for judgment are pertinent to measurement of the evaluation abilities; and (*c*) whether there are any new restrictions to be placed on the concept. The answers to these issues rest upon the kinds of tests that serve to measure their evaluation factors well and those that do not, when all have been hypothesized to do so.

Kinds of judgment. Much was said from one place to another in this report about the two classes of tests: sensitivity versus estimation. More fully spelled out, these terms mean sensitivity to error or discrepancy on the one hand and, on the other, judgment of relative nearness of a number of items of information (for any kind of product) to a kind of model item of information on the same continuum. In more operational terms, the contrast may be stated in terms of "absolute" versus "relative" kinds of judgments, as in psychophysics.

It should be abundantly clear, from the way in which both sensitivity and estimation types of tests are commonly related significantly to the same factors, that both kinds of judgments apply. If we compare evaluation tests from the two categories that clearly involve absolute versus relative judgments, however, we find that some factors tend to be more strongly defined by tests of absolute judgments or sensitivity, while others tend to be defined by tests of relative judgment or estimation. It is possible that further research will emphasize such between-factor differences, but at present it is safe to assume that each factor can be measured by both types of tests.

Of the 14 leading univocal tests for the 12 obtained evaluation factors, 11 were tests of relative judgment or estimation and 3 were of absolute judgment or sensitivity. It should be noted, however, that the majority of the experimental tests were of the former type, and therefore chance alone would favor such a result for relative-judgment tests.

Kinds of criteria. The most common criteria employed in evaluation tests have been identity versus nonidentity (Perceptual Speed) and consistency versus inconsistency (Logical Reasoning). Other kinds of criteria have been mentioned in connection with various tests and for different factors. Some examples are: fitness for class membership (Letter "U" and Double Descriptions); relative familiarity (Familiar Letter Combinations); relative similarity (Series Relations and Verbal Analogies III); conformity to principles (Correct Number Series, Correct Letter Orders, and Word Systems); relative probability (Decoding and Unlikely Things); and solvability of problems (Letter Problems). Such a variety of criteria can possibly be brought under more abstract and more general criterion categories, since such terms as "identity," "similarity," and "conformity" suggest that the criteria tend to be logical in nature and that they represent continua of one kind or another.

The scope of evaluation. The scope of processes under the heading of evaluation

is indicated somewhat by the variety of criteria that may be involved. The discussion of this topic above revealed something of the apparent variety of criteria that is represented in the experimental tests. But it was suggested that such criteria may be limited to the general logical category, which would rule out of consideration criteria involving esthetic and ethical values. There is no doubt that such values exist and such areas of judgment call for evaluative operations. At the present, they do not seem to fit into the structure-of-intellect model. Perhaps they call for two complete additional sets of evaluation abilities or processes, whether parallel to the present theoretical set or not. Another possibility is that esthetic judgments may be applicable in the decision processes concerning figural information, and ethical or moral ones concerning behavioral information. Such an extension of the concept of evaluation awaits further investigation.

As to the definition of evaluation itself, the kind of restriction just discussed suggests that it is going too far to say that evaluation is a matter of reaching decisions regarding goal satisfaction. Kinds of goals are much too numerous, and satisfaction in terms of logical criteria cannot cover all cases. In defining the restricted kinds of evaluation represented in the structure of intellect, it would seem desirable to eliminate reference to "goal satisfaction."

As a general impression, from consideration of the experimental tests and their factors in this study, the importance of an act of comparison seems to stand out. The observation was also made by Hertzka et al. (1954) that the core of the definition of evaluation is the concept of comparison. The following current definition of evaluation can be suggested: Evaluation is a process of comparing information with known information according to logical criteria, reaching a decision concerning criterion satisfaction.

Recommended Tests for the Evaluation Factors

Although many of the experimental tests were disappointing, loading on too many factors to be considered univocal, or not loading on factors for which they were designed and in terms of which they could be understood, several performed in a manner that provides some confidence in recommending them tentatively as the best available measures of their factors. The recommended tests appeared to be fairly univocal and, in other respects, reasonable measures of their factors. The recommended measures for the experimental evaluation factors are:

ESU	Symbol Identities
	Letter "U"
ESC	Best Number Class
	Sign Changes II
ESR	Symbol Manipulation
	Related Words I
ESS	Way-Out Numbers
	Series Relations
EST	Jumbled Words
ESI	Abbreviations
EMU	Double Descriptions
	Sentensense
EMC	Best Word Class
	Class Name Selection
EMR	Verbal Analogies III
	Best Trend Name
EMS	Word Systems
	Unlikely Things
EMT	Useful Changes
EMI	Complete Thoughts
	Sentence Selection

Summary

Two studies were designed to test the implications of, and extend the empirical foundations underlying the structure-of-intellect model. The studies attempted to identify basic traits with respect to which individuals differ from one another in evaluative performances. The 12 hypothesized abilities of symbolic and semantic evaluation were selected for investigation. The major objective was to determine whether such distinguishable abilities could be demonstrated; distinguishable from one another and also from other intellectual abilities. A secondary objective was the determination of what mental processes are evaluative.

The 12 hypothesized evaluation factors are: six symbolic factors investigated in

the first study—evaluation of symbolic units (ESU); evaluation of symbolic classes (ESC); evaluation of symbolic relations (ESR); evaluation of symbolic systems (ESS); evaluation of symbolic transformations (EST); and evaluation of symbolic implications (ESI); and six semantic factors investigated in the second study—evaluation of semantic units (EMU); evaluation of semantic classes (EMC); evaluation of semantic relations (EMR); evaluation of semantic systems (EMS); evaluation of semantic transformations (EMT); and evaluation of semantic implications (EMI).

In order to demonstrate the distinctness of the hypothesized factors from those already known, 19 reference factors, previously confirmed, were analyzed as experimental controls. The reference factors for the symbolic study included all of the known cognition and convergent-production factors concerned with symbolic information, verbal comprehension, numerical facility, a symbolic flexibility factor, and perceptual speed. In the semantic study, the reference factors included all the known cognition factors concerned with the semantic information, also semantic elaboration, and redefinition. From the list of reference factors it can be seen that special attention was paid to differentiating the evaluation factors from parallel factors of cognition.

Two different 8-hour test batteries were constructed to accomplish the experimental objectives. A total of 50 tests (25 evaluation tests and 25 markers) were administered to 225 high-school seniors for the symbolic analysis and 41 measures (22 evaluation tests and 19 markers) were administered to 202 high-school juniors for the semantic analysis. Principal-axes factors were obtained, 18 for the symbolic analysis and 14 for the semantic analysis, and were rotated both graphically and analytically, observing the criteria of simple structure, positive manifold, and psychological meaningfulness.

All the hypothesized factors emerged, defined largely by the tests designed to measure them. The reference factors were also successfully isolated, indicating generally good factorial invariance for the marker tests.

Among the six symbolic-evaluation factors, ESU was clearly defined as a factor that could be described nontechnically as clerical speed and accuracy, the ability to judge rapidly symbolic material in terms of identity or error. The ESC factor was least clear-cut, but it involved sensitivity to class properties. ESR was clearly isolated as the ability to make choices among symbolic relationships on the bases of identity and consistency. The ESS factor appeared to involve the estimation of similarity among and values within symbolic series. The clear isolation of EST defined it as the ability of sensitivity to the fulfillment of criteria by rearrangements and substitutions of letters within words. The ability to judge the consistency or probability of implications from symbolic material was represented by the ESI factor.

Of the six hypothesized factors of semantic evaluation, EMU, EMC, and EMR are essentially new factors, well supported by the results and parallel in nature to their analogous symbolic factors. EMR replaces the factor formerly known as "logical evaluation," erroneously allocated to the EMR cell of the structure of intellect. Factor EMS was essentially a verification of the previously recognized factor of "experiential evaluation," but the new EMS factor is somewhat broader. A new factor replaces "sensitivity to problems" with a better claim to the cell EMI. The implications factor involves decision making regarding reasonableness of relatively remote consequences or expectancies.

Relatively clear separation among abilities according to operation, content, and product was interpreted as substantially contributing to the value of the unified theoretical model of the structure of intellect as a hypothetico-deductive theory for generating predictions concerning individual differences in intellectual functioning.

REFERENCES

BECHTOLDT, H. P. Factorial investigation of the perceptual speed factor. *American Psychologist*, 1947, 2, 304–305.

BERGER, R. M., GUILFORD, J. P., & CHRISTENSEN, P. R. A factor-analytic study of planning abilities. *Psychological Monograph*, 1957, 71, (6, Whole No. 435).

CATTELL, R. B. A universal index for psychological factors. *Advanced publication number 3, labora-*

tory of personality assessment as group behavior. Urbana: University of Illinois, 1953.

CLIFF, N. Orthogonal rotations to congruence. *American Psychologist,* 1964, **19,** 582. (Abstract)

COOMBS, C. H. A factorial study of number ability. *Psychometrika,* 1941, **6,** 161–189.

DEWEY, J. *How we think.* Boston: Health, 1910.

FRENCH, J. W. The description of aptitude and achievement tests in terms of rotated factors. *Psychometric Monograph,* No. 5, 1951.

FRENCH, J. W., EKSTROM, R. B., & PRICE, L. A. *Manual for kit of reference tests for cognitive factors.* Princeton: Educational Testing Service, 1963.

FRICK, J. W., GUILFORD, J. P., CHRISTENSEN, P. R., & MERRIFIELD, P. R. A factor-analytic study of flexibility in thinking. *Educational and Psychological Measurement,* 1959, **19,** 469–496.

GREEN, R. F., GUILFORD, J. P., CHRISTENSEN, P. R., & COMREY, A. L. A factor-analytic study of reasoning abilities. *Psychometrika,* 1953, **18,** 135–160.

GUILFORD, J. P. Three faces of intellect. *American Psychologist,* 1959, **14,** 469–479.

GUILFORD, J. P. *Fundamental statistics in psychology and education.* (4th ed.) New York: McGraw-Hill, 1965.

GUILFORD, J. P., GREEN, R. F., CHRISTENSEN, P. R., HERTZKA, A. F., & KETTNER, N. W. A factor-analytic study of Navy reasoning tests with the Air Force Aircrew Classification Battery. *Educational and Psychological Measurement,* 1954, **14,** 301–325.

GUILFORD, J. P., & HOEPFNER, R. Current summary of structure-of-intellect factors and suggested tests. *Report from the Psychological Laboratory,* Number 30. Los Angeles: University of Southern California, 1963.

GUILFORD, J. P., HOEPFNER, R., & PETERSEN, H. Predicting achievement in ninth-grade mathematics from measures of intellectual-aptitude factors. *Educational and Psychological Measurement,* 1965, **25,** 659–682.

GUILFORD, J. P., & LACEY, J. I. (Eds.) *Printed classification tests.* Army Air Force Aviation Psychology Research Reports, Report Number 5. Washington, D.C.: Government Printing Office, 1947.

GUILFORD, J. P., & MERRIFIELD, P. R. The structure of intellect model: Its uses and implications. *Report from the Psychological Laboratory,* Number 24. Los Angeles: University of Southern California, 1960.

GUILFORD, J. P., MERRIFIELD, P. R., CHRISTENSEN, P. R., & FRICK, J. W. Some new symbolic factors of cognition and convergent production. *Educational and Psychological Measurement,* 1961, **21,** 515–541.

GUILFORD, J. P., MERRIFIELD, P. R., & COX, A. B. Creative thinking in children at the junior high school levels. *Report from the Psychological Laboratory,* Number 26. Los Angeles: University of Southern California, 1961.

GULLIKSEN, H. *Theory of mental tests.* New York: Wiley, 1950.

HERTZKA, A. F., GUILFORD, J. P., CHRISTENSEN, P. R., & BERGER, R. M. A factor-analytic study of evaluative abilities. *Educational and Psychological Measurement,* 1954, **14,** 581–597.

HOEPFNER, R., & GUILFORD, J. P. Figural, symbolic, and semantic factors of creative potential in ninth-grade students. *Report from the Psychological Laboratory,* Number 35. Los Angeles: University of Southern California, 1965.

HOEPFNER, R., GUILFORD, J. P., & MERRIFIELD, P. R. A factor analysis of the symbolic-evaluation abilities. *Report from the Psychological Laboratory* Number 33. Los Angeles: University of Southern California, 1964.

JOHNSON, D. M. *The psychology of thought and judgment.* New York: Harper, 1955.

KETTNER, N. W., GUILFORD, J. P., & CHRISTENSEN, P. R. A factor-analytic study of the factor called general reasoning. *Educational and Psychological Measurement,* 1956, **16,** 438–453.

KETTNER, N. W., GUILFORD, J. P., & CHRISTENSEN, P. R. A factor-analytic study across the domains of reasoning, creativity, and evaluation. *Psychological Monograph, 1959,* **73,** (9, Whole No. 479). (a)

KETTNER, N. W., GUILFORD, J. P., & CHRISTENSEN, P. R. The relation of certain thinking factors in training criteria in the U.S. Coast Guard Academy. *Educational and Psychological Measurement,* 1959, **19,** 381–394. (b)

MARKS, A., GUILFORD, J. P., & MERRIFIELD, P. R. A study of military leadership in relation to selected intellectual factors. *Report from the Psychological Laboratory,* Number 21. Los Angeles: University of Southern California, 1959.

MERRIFIELD, P. R., GUILFORD, J. P., CHRISTENSEN, P. R., & FRICK, J. W. The role of intellectual factors in problem solving. *Psychological Monograph,* 1962, 76, (10, Whole No. 529).

NIHIRA, K., GUILFORD, J. P., HOEPFNER, R., & MERRIFIELD, P. R. A factor analysis of the semantic-evaluation abilities. *Report from the Psychological Loboratory,* Number 32. Los Angeles: University of Southern California, 1964.

PEMBERTON, C. L. The closure factors related to other cognitive processes. *Psychometrika,* 1952, **17,** 267–288.

THURSTONE, L. L. Primary mental abilities. *Psychometric Monograph,* No. 1, 1938. (a)

THURSTONE, L. L. The perceptual factor. *Psychometrika,* 1938, **3,** 1–17. (b)

THURSTONE, L. L. A factorial study of perception. *Psychometric Monograph,* No. 4, 1944.

UNDERWOOD, B. J., & SCHULTZ, R. W. *Meaningfulness and verbal learning.* Philadelphia: Lippincott, 1960.

WILSON, R. C., GUILFORD, J. P., CHRISTENSEN, P. R., & LEWIS, D. J. A factor-analytic study of creative-thinking abilities. *Psychometrika,* 1954, **19,** 297–311.

(Received April 18, 1966)

Vol. 80, No. 17 Whole No. 625, 1966

Psychological Monographs: General and Applied

THE ASSESSMENT CENTER IN THE MEASUREMENT OF POTENTIAL FOR BUSINESS MANAGEMENT

DOUGLAS W. BRAY AND DONALD L. GRANT[1]

American Telephone and Telegraph Company

The assessment process in the Bell System's Management Progress Study is described, and the results of several analyses of the process are presented. Included are studies of assessment staff evaluations, contributions to the process of selected techniques, and relationships of assessment data to subsequent progress in management. The results, based on 355 young managers, indicate that the evaluations by the assessment staffs were influenced considerably by their overall judgments of the men assessed but also made many intraindividual discriminations. The results also show that all of the techniques studied made at least some contribution to the judgments of the assessors. Situational methods (group exercises and In-Basket) had considerable influence; paper-and-pencil ability tests had somewhat less influence; personality questionnaires were given the least weight. (Projective methods and interviews were not included in the analyses but are being studied.) The relationships between assessor judgments and subsequent progress in management, though covering only a relatively short time period, indicate that the assessors' predictions were quite accurate. The results also show that a complex of personal characteristics is more predictive of progress than any single characteristic. Some of the characteristics, however, appear to have higher relationships to progress than do others. Of the techniques studied the situational methods and paper-and-pencil ability tests are more predictive of progress than the personality questionnaires.

The assessment center method of evaluating individual characteristics and potential, although not widely studied or used because of its expense and complexity, has continued to command interest. Early applications of the method occurred primarily in military or academic contexts, but there is presently a growing experimentation with assessment centers in American business as a method of evaluating managerial ability.

The Bell System's Management Progress Study (Bray, 1964) offers a unique opportunity to study the assessment process. The Study, which began in 1956, is a longitudinal study of the development of young men in a business management environment. Assessment per se is only one of the several research methods being used.

The uniqueness of the Study from the standpoint of studying the assessment process arises from several aspects of the Study design:

1. There is no contamination by the assessment results of the subsequent criterion data. Along with all other information collected on the 422 subjects of the Study, the assessment data are being held in strict confidence. Thus the judgments of the assessment staff have had no influence on the careers of the men being studied.

2. The subjects of the Study have been or will be reassessed. A means thereby exists for taking into account the effects of growth on assessed characteristics.

3. Because of the longitudinal nature of the Study and of the vast amount of data being accumulated, there is very little limit to the number of kinds of analyses involving the assessment data which can be made. As a result, the interrelationships of the assessment data with a variety of criteria can and are being investigated.

Nature of Assessment

The origin of the use of multiple assessment procedures on a large scale is credited

[1] The senior author is responsible for the design of the Management Progress Study and the assessment center method used in the Study; the junior author planned and carried out all the analyses included in this report.

to German military psychologists (OSS, 1948). The British adapted the procedures to the screening of officer candidates, and the United States Office of Strategic Services (OSS) took over the approach from the British during World War II (OSS, 1948). Since that time several studies of various applications of these procedures have been reported in the literature (see especially Taft, 1959, and Cronbach, 1960, for summaries of such studies).

In general, the procedures employed have involved the use of multiple methods for obtaining information on individuals, standardization of these methods and those for making inferences from them, and the use of several assessors whose judgments are pooled in arriving at evaluations of the persons assessed. As might be expected, many variations in methodology have been reported.

A major contribution of the multiple assessment approach has been the use of situational tests or exercises. Though not restricted to multiple assessment, the application of situational techniques to assessment has been featured in such programs as that of the OSS.

Situational methods offer the potential of adding to the scope of human characteristics which can be evaluated. Though much more expensive and time consuming to administer than paper-and-pencil tests and questionnaires, the need to find ways of evaluating characteristics not covered by the latter is sufficient to warrant extensive experimentation with relatively elaborate techniques.

Assessment procedures also contrast with psychometric ones in the way the resulting data are combined. Psychometric approaches depend on mathematical methods for accomplishing this purpose whereas assessment approaches combine the data judgmentally.

Experience with Assessment

Taft (1959) has pointed out that the reasons for using multiple assessment approaches have varied (e.g., personality research, selection, validation of techniques). As a consequence, the foci of research using such procedures have differed. He noted, however, that "All assessment programs involve studies of the link between two or more pieces of behavior.... Some of this behavior is known as assessment behavior and some as criterion behavior [p. 377]."

The majority of studies reporting use of multiple assessment procedures have focused on prediction. Many of these studies, particularly several of the earlier ones, reported results which raised serious questions regarding the "predictive validity" of such procedures. Other studies, however, have tended to support the value of using assessment approaches for predictive purposes.

Among the former are those by the OSS (1948), the study by the Veterans Administration of clinical psychologists (Kelly & Fiske, 1951; Kelly & Goldberg, 1959), the Menninger School of Psychiatry study (Holt & Luborsky, 1958), and the study of Air Force officers (MacKinnon et al., 1958). Studies reporting relatively high predictive validities for the assessment procedures used include two studies cited by Cronbach (1960), that is, those of British civil service candidates (Vernon, 1950) and of American OCS applicants (Holmen, Katler, Jones, & Richardson, 1956). In addition, two relatively recent studies, one of Scandanavian airline pilots (Trankell, 1959), and another of managerial personnel (Albrecht, Glaser, & Marks, 1964) report relatively high correlations between assessment results and subsequent measures of performance.

Though no firm conclusions regarding the predictive validities of multiple assessment procedures can be drawn from the rather mixed findings of published research, it does appear clear that the more accurate predictions were obtained where the performance to be predicted was clearly defined, the assessment results did not restrict the range of subsequent criterion performance, and the criterion measures employed were not not limited by low reliability and questionable validity. None of the published studies, incidentally, report completely invalid results, though in some the correlations with performance criteria are

disappointingly low. Furthermore, in such studies as that of clinical psychologists (Kelly & Fiske, 1951) the paper-and-pencil tests used predict subsequent performance as well as do the assessment ratings.

Theoretical and Methodological Considerations

As has been noted (Albrecht et al., 1964) assessment procedures have been applied without the benefit of much prior developmental effort, either theoretical or empirical. The OSS (1948) did follow a rather well-developed rationale in applying the procedures, though it is evident from their report that considerable trial and error accompanied application. Many modifications in the procedures were made as the program developed.

Subsequent studies have contributed little to the formulation and testing of assessment principles. After reviewing the pertinent literature Taft (1959) discusses many of the issues involved, including the different strategies used in predicting criteria performance. Stern, Stein, and Bloom (1956) have published the most thorough discussion, illustrated by small-sample studies, of assessment "models." Four alternative approaches are described along with considerations of the advantages and disadvantages of each.

Though no firm set of principles have emerged from such discussions, certain aspects of assessment have been highlighted which warrant consideration, even though brief.

Prior analysis. Much emphasis has been placed on the need for a thorough study of the total situation for which the assessment is being made. Included are such factors as behavioral requirements, environmental influences, functional roles, and value judgments of "significant others" (Stern, Stein, & Bloom, 1956). From such an analysis are derived the variables for which assessment staff judgments are to be made.

Assessed characteristics. Based on the prior analysis, the characteristics assessed (usually including an "overall" evaluation) must be defined in behavioral terms so as to facilitate appropriate judgments by the assessment staff. No firm set of principles for determining the number of characteristics to be assessed nor for assuring adequate definitions have been advanced. In practice the number and nature of such variables have varied widely. Principles of rating, such as developed by Wherry (1952), are pertinent, however, to making such decisions.

Techniques. Methods for obtaining relevant information on the persons to be assessed are selected or developed in accord with the characteristics for which judgments are to be made. Again, no firm set of principles for making these decisions have been advanced. Multiple methods have been favored by many practitioners. In practice, a variety of methods (including interviews, situational tests, paper-and-pencil tests, biographical questionnaires, and projectives) have been used. Little information on the relative values to the assessment process of the various methods has been reported. The number of methods used also has varied widely.

Staffing. Presumably, the success of assessment depends considerably on the competence of the assessors. If so, the size and organization of the staff, their selection, the quality of training given them, and their supervision should influence the results obtained. Professionals have tended to favor professionally trained persons for assessment activities, though in one of the studies cited by Cronbach (1960) an "amateur" staff provided quite valid predictions. Again, pertinent information leading to a set of principles is needed.

The assessed. In practice, the numbers of persons assessed at a particular time have varied considerably. Logistical requirements obviously have influenced decisions regarding this aspect of assessment. Presumably, however, some optimal ratio of assessees to staff and to the number and nature of techniques may exist, but have not been elaborated.

Evaluating the assessees. The final step in the assessment process is that of evaluating the assessees. The entire assessment staff, or selected individuals therefrom, review the evidence and rate each person on

each of the assessment variables. Again, there have been many variations in the rating procedures used. Little attention has been given such factors as the number of raters, the time lag between observation and rating, the number of assessees rated at a time, and the mechanics for pooling individual ratings. Principles of rating such as developed by Wherry (1952), previously referred to, have pertinence to this crucial step in the assessment process.

Evaluating results of assessment. Determining the "validity" of procedures has been a major concern of practitioners of this art. Much attention has been given to "predictive" validity, very little to "construct" validity. Problems in research design have been discussed at length by several investigators (e.g., Albrecht et al., 1964; OSS, 1948). Criterion problems have proven as thorny for these investigators as they have for the psychometrician. Furthermore, where prior screening has been effective and/or the assessment results have influenced personnel decisions allowance for the consequent restrictions on range of subsequent performance has been inadequate. It seems reasonable to believe that an accurate evaluation of assessment results requires appropriate criterion measures, a representative range of criterion behavior, including sufficient time following assessment to permit the development of relevant criterion behavior, and adequate controls over other factors which may introduce irrelevance in the criterion measures.

As will be seen in the ensuing section of this report, many of the considerations discussed above were taken into account in designing the assessment phase of the Study. The "model" employed was the one developed by the OSS (1948), modified to fit the requirements of the Study.

Because the focus of the Study is not on assessment per se, no attempt has been made to test out alternative approaches to assessment. The assessment procedures were modified somewhat, however, in light of initial experience with them. Following initial changes the remainder of the assessment work was carried out with standardized procedures.

This report is directed at presenting the early findings on assessment procedures deriving from the Management Progress Study. Future reports will present additional results. Covered in this report are:

1. Descriptions of the assessment procedures.
2. Analyses of the assessment staff evaluations.
3. Contributions to the assessment process of selected techniques.
4. Relationships of assessment data to progress in management over relatively short time periods.

ASSESSMENT PROCEDURES

The subjects of the Management Progress Study were assessed at the time of their inclusion in the Study. The purpose of the assessment was to measure personal characteristics hypothesized to be of importance either in developmental change in early adulthood or success in business management.

The sample of 422 men were at the time of assessment employees of six Bell System telephone companies. Approximately two-thirds of the sample were college graduates who were assessed soon after employment. The remaining third had been employed initially for nonmanagement positions and had advanced into management relatively early in their careers. These men were not college graduates when employed; a few had since earned degrees by part-time and evening study.

The first step in designing the assessment procedures was that of selecting and defining those characteristics to be assessed. In doing this a thorough review of the literature was supplemented by securing the judgments of experienced Bell System personnel men as to the qualities they believed to be most important in success in the business. The many characteristics which this process produced were finally reduced to a list of 25 qualities. Techniques (described below) were then selected or developed to reveal these variables.

The assessment staffs consisted primarily of professionally trained persons, though as the Study progressed a few telephone company managers served on the staffs. None of the latter, incidentally, came from the company of the men being assessed.

The subjects spent 3½ days at the assessment center in groups of 12. Immediately following, the assessment staff conducted extensive discussions of each participant and rated each on the 25 characteristics. At a later date a narrative summary of each man's performance was prepared.

Assessment of the subjects was spread over several summers. The first ones assessed, all the college graduates employed by one telephone

company in that year, were assessed during the summer of 1956. The last to be assessed were processed during the summer of 1960. Modifications in the methods used were made subsequent to the 1956 assessment. Thereafter, the methods remained standard.

Techniques

The methods used for collecting information on the personal characteristics of the participants are representative of those used generally in assessment activities. A listing of the techniques with a brief description of each follows:

Interview. A 2-hour interview with each man directed at obtaining insights into his personal development up to that time, work objectives, attitudes toward the Bell System, social values, scope of interests, interpersonal relationships, idiosyncrasies, etc.

In-Basket. A set of materials which a telephone company manager might expect to find in his in-basket. The items, 25 altogether, range from telephone messages to detailed reports. In addition, examinee was furnished with such necessary materials as a copy of the union contract, organization chart, and stationery. He was given 3 hours in which to review the materials and take appropriate action on each item (by writing letters, memos, and notes to himself). Following completion of the "basket" he was interviewed concerning his approach to the task, his reasons for taking the actions indicated, and his views of his superiors, peers, and subordinates (as inferred from the materials).

Manufacturing Problem (made available by John Hemphill of the Educational Testing Service). A small-business game wherein the participants assumed the roles of partners in an enterprise manufacturing toys for the Christmas trade. The participants were required to buy parts and sell finished products under varying market conditions, to maintain inventories, and to manufacture the toys.

Group Discussion. Also a leaderless group situation, focused around a management personnel function. Participants were instructed to assume the roles of managers, each having a foreman reporting to him considered capable of promotion. Participants were required to discuss the merits and liabilities of their hypothetical foremen and to reach a group decision regarding their relative promotabilities.

Projectives. (*a*) Rotter Incomplete Sentences Blank (published by the Psychological Corporation). (*b*) Bell Incomplete Sentences Test (by Walter Katkovsky and Vaughn Crandall with the advice and assistance of Julian Rotter). (*c*) Thematic Apperception Test (published by Harvard University Press). Six of the cards from this test were administered.

Paper-and-pencil tests and questionnaires. (*a*) School and College Ability Test, Form 1 (published by the Cooperative Test Division of The Educational Testing Service). (*b*) A Test of Critical Thinking in Social Science (American Council on Education, 1951). Now out of print, this test was designed to measure several aspects of critical thinking, including the ability to define problems, select pertinent information, recognize unstated assumptions, evaluate hypotheses, and make valid inferences. (*c*) Contemporary Affairs Test (formerly published by the Cooperative Test Division of the Educational Testing Service). Annually since 1956 the Personnel Research Section of AT & T has developed, following the Educational Testing Service format, its own version of this test. (*d*) Edwards Personal Preference Schedule (published by the Psychological Corporation). (*e*) The Guilford-Martin Inventory of Factors GAMIN (published by the Sheridan Supply Company). (*f*) Opinion Questionnaire, Form B. Unpublished, this questionnaire was adapted from Bass (1955) and yields three scores —authoritarianism (A), acquiesence (a), and negativism (n). (*g*) Survey of Attitudes Toward Life. Unpublished, this questionnaire, made available to the Bell System by Irving Sarnoff of New York University, is designed to reflect a person's attitudes toward making money and advancing himself.

Miscellaneous. (*a*) Personal history questionnaire. (*b*) Short autobiographical essay. (*c*) *Q* sort (70 items, self-descriptive).

Administration and Reporting

All of the assessment techniques were administered according to standard instructions by the staff member or members responsible for each method. Naturally, the ways in which this was accomplished varied according to the nature of the technique. Thus all interviews were conducted by individual staff members with individual assessees. All tests, questionnaires, and situational exercises were administered to groups of participants. The two group problems involved six participants at a time. Two staff members observed each group problem, recorded their impressions of each participant, and independently evaluated the performance of each.

For each method one or more of the staff prepared written reports. The interviewers dictated, as nearly verbatim as possible, reports of their interviews. One staff member reviewed each completed In-Basket, along with the accompanying interviewer's report on handling the "basket," and prepared reports describing how each man dealt with the materials in the exercise along with evaluating his effectiveness in so doing. Similarly, one of the observers for each group problem prepared reports describing and evaluating individual performance in these exercises, including ratings and rankings by both peers and observers. A clinically trained psychologist reviewed the projective protocols and prepared individual reports. The paper-and-pencil tests were scored and summarized.

Rating Variables

The personal characteristics selected for evaluation reflect varied aspects of what could be referred to as "criterion" performance, broadly conceived. Some are directly related to managerial functions (e.g., organizing, planning, decision making, problem solving). Others refer to interpersonal relationships and influence (e.g., communications skills, personal impression, sensitivity, dependence on others). Still others relate to general abilities (e.g., intellectual ability, adaptability).

Motives, values, and attitudes are covered by several of the variables. Included are attitudes toward the importance of work and toward working for a large company. Social attitudes are included as are desires for advancement and security. Personal goals, self-evaluations, and expectations also were evaluated.

Staff Evaluations

Immediately following the 3½ days of collecting information on the subjects, the assessment staff, consisting usually of nine persons, assembled, reviewed, and discussed the results. Each man assessed was evaluated separately, 1 to 1½ hours being required per man.

A typical evaluation consisted of first reading the man's short autobiographical essay. An interviewer then read a summary of his interview with the man. Reports on the In-Basket, group exercises, paper-and-pencil test scores, and projectives followed, concluding with a reading of the *Q*-sort items the man selected as "most" and "least" like him.

Following presentation of the reports each staff member independently rated the man on each of the 25 characteristics (from 1 [low] to 5 [high]). Each of the variables then was reviewed. Where differences of opinion occurred the evidence was discussed and staff members permitted, though not required, to adjust their ratings.

After the variables had been rated, the staff evaluated the man's potential as a management person in the Bell System. Separate judgments were recorded, independently, regarding the man's likelihood of remaining in the Bell System and, assuming that he would remain, of achieving middle management in 10 or less years. In addition, the staff noted their judgments as to whether the man "should" advance to middle management. Again, the ratings of potential were discussed and adjusted where staff members wished to do so.

For analysis purposes all of the data thus collected were filed, using code numbers to protect anonymity, at the Fels Research Institute at Antioch College. So as to obtain a consensus rating on each variable the variable ratings were averaged. The ratings of potential, however, were trichotomized in order to reflect the extent of agreement by the staff (No, ?, Yes). In the analyses to be presented these consensus ratings were used.

Analyses of the Staff Evaluations

To date, three kinds of studies of the assessment process have been undertaken, namely:

1. Studies of the interrelations between the rating variables and their relationships to the overall predictions of the assessment staff.
2. Studies of the assessment techniques with particular reference to their contributions to the assessment process.
3. Studies of relationships between staff evaluations and such behavioral criteria as survival in the business and progress in management.

Analyses of assessment staff evaluations have been made by intercorrelating the ratings on the 25 variables, correlating the ratings with the overall predictions of the staff, and factoring the intercorrelations between the variables. The results of the latter are presented in the following pages.

The reasons for factoring the rating variables are twofold. In the first place it was expected that the factorial results would result in clarifying the nature of the judgments made. Mere inspection of the 25 variables undoubtedly would suggest many of the underlying constructs employed in making the judgments. Factorial results, however, should permit greater precision in interpreting the variables and also make it possible to avoid misinterpretations.

Factorial results also can provide a useful method for organizing ratings for subsequent analyses. Composite scores based on factors are presumably more reliable than those based on individual variables and, being fewer in number, can be more efficiently utilized.

Method

The entire sample of 422 men was divided according to educational background at time of employment. The ratings for all who were not college graduates at time of employment (N = 148) and those for 207 of the college graduates were used in the analyses. (Because of revisions in the assessment methods, ratings for 67 college graduates from the first telephone company in-

volved in the study were excluded from the analyses.) Separate analyses were made for each sample.

For each sample the rating variables and the staff prediction regarding the likelihood of progressing to middle management in 10 or less years were intercorrelated. The product-moment correlations computed are presented in Appendix A. The resulting matrices were then factored by a method developed by Wherry (1959) for determining a hierarchical factor solution. Though the method is a general one it appears particularly relevant to rating data, which characteristically are influenced by "halo effects." The higher-order factors obtained presumably reflect the latter, whereas the lower factors reflect more specific judgments by the raters.

Results

The factorial solutions yielded 11 factors for the college graduate sample and 8 factors for the nongraduate sample. The loadings, paired for comparable factors, are shown in Appendix B. That the results account for relatively large shares of the total variances of the ratings for both samples is indicated by the magnitude of the communalities. The average communality for the college group is .64, that for the noncollege being .57.

Both solutions yielded higher-order factors. For the noncollege sample a single "general" factor was obtained (Factor I). This factor separated into three higher-order factors for the college sample (a third-order general factor and two second-order subgeneral factors, Factors II and III).

Factor I has similar patterns of loadings in both samples. For that matter, Factor I could be described as reflecting the assessment staff's "model" for managerial potential (the loadings of the staff predictions being highest on this factor). In general, a man rated high on this factor was seen as effective in organizing, planning, and decision making, likely to solve a management problem in a novel way, skillful in dealing with others, resistant to stress, above average in intellectual competence, able to communicate orally, energetic, and perceptive of the behavior of others. Motivationally, he was seen as desirous of advancing in the management hierarchy, having high standards of work performance, not particularly concerned with job security, and relatively independent of the approval of his peers and superiors.

Factor II (college sample only) also is indicative of managerial potential (having nearly as high a loading on the staff prediction as does Factor I). Those achieving high ratings on this factor are similar in many characteristics to the more highly rated on Factor I. They differ chiefly in the motivational areas, being less likely to value advancement in the management hierarchy and more likely to value high performance standards in the work itself. They also may or may not be dependent on the approval of their peers and superiors and may or may not value a secure job. They are somewhat more likely to be skilled in dealing with others though less likely to stand up under stress.

Factor III (college sample only) has its highest loadings on several motivational variables. Those evaluated high on this factor are characterized by willingness to postpone rewards, particularly advancement in the organization, along with wanting a secure job. They seek the approval of their peers and superiors and are likely to incorporate company values. Because such needs appear indicative of passivity (avoiding competition and risk taking) on the one hand and dependency (desiring support from others) on the other, this factor is named "passive dependency."

It is not self-evident from the data as to why a single higher-order factor resulted from the analysis of the correlations from the noncollege sample whereas three such factors were generated by the analysis of the college sample matrix. One can speculate that the college sample was more homogeneous with respect to the characteristics evaluated, though further study would be required to verify this speculation.

From the analyses seven first-order factors were determined for the college sample and six for the noncollege sample. These factors reflect the more specific judgments of the assessment staff. Summary descriptions of each factor, including variables with the highest loadings, follow:

Factor	Sample	Variables	Name
IV	Both	Organizing and planning Decision making	Administrative skills
V	Both	Human relations skills Behavior flexibility Personal impact	Interpersonal skills
VI	Both	Tolerance of uncertainty Resistance to stress	Control of feelings
VII	Both	Scholastic aptitude Range of interests	Intellectual ability
VIII	Both	Primacy of work Inner work standards	Work-oriented motivation
IX	Both	Ability to delay gratification Need for security Need for advancement (negative)	Passivity
X	Both	Need for superior approval Need for peer approval Goal flexibility	Dependency
XI	College	Social objectivity Range of interests Need for security (negative) Need for superior approval (negative) Bell System value orientation (negative)	Nonconformity

The relative importance of the judgments by the assessment staff of general effectiveness is apparent from the analyses. Factors I and II account for 30% of the average total variance (47% of the accounted-for variance) in the college sample while Factor I accounts for 26% of the average total variance (45% of the accounted-for variance) in the noncollege sample. In brief, as might be expected, the raters were influenced by overall impressions of the men assessed. The method of making the ratings, across the variables, might be expected, of course, to enhance any halo tendencies that were present.

Further evidence regarding the weight given overall impact is obtained from inspection of the loadings on the staff predictions of potential (will reach middle management in 10 or less years). It is apparent that the general factors (I and II) are the most heavily weighted on this variable. For the college sample these factors account for 61% of the total variance (73% of the accounted-for variance) of the staff predictions while for the noncollege sample Factor I accounts for 42% of the total variance (64% of the accounted-for variance). Much lesser weights were obtained for interpersonal skills, intellectual ability, and nonconformity (college sample) and intellectual ability and passivity (noncollege sample), while the remaining lower-order factors have practically zero weights. Thus in making predictions of progress in the management hierarchy the assessment staffs evidently were primarily influenced by their overall judgments and secondarily by more specific evaluations.

Despite such tendencies, however, the assessment staffs were able to make many discriminations on more specific variables. Eight first-order factors for the college sample and seven for the noncollege sample are relevant evidence as is the fact that for both samples over half of the average accounted-for variance can be ascribed to the more specific factors.

Though there are some discrepancies, the consistency in the factor structure from sample to sample is quite high. Because the assessment process was the same for both samples it is hardly surprising, despite differences in the educational backgrounds of the men assessed, that the raters were influenced by similar considerations.

The nature of the more specific factors is of interest. Three reflect abilities (administrative skills, interpersonal skills and intellectual ability) whereas five reflect temperament and motivation (control of feelings, work-oriented motivation, passivity, dependency, and nonconformity). The analyses makes it apparent that the methods

used in observing the men and the variables employed in evaluating performance permit consideration of a wide range of characteristics.

The factorial results also help to clarify the constructs used by the staff evaluators. Mere inspection of the variables would suggest such factors as administrative skills, intellectual ability, and control of feelings. Whether factors like passivity and nonconformity would result from such an inspection is doubtful. Furthermore, some of the variables would have been difficult to evaluate.

The variable "goal flexibility" can be used to exemplify this point. One might not hypothesize that ratings on this variable reflect dependency needs. The factorial results, however, suggest such an interpretation. It seems reasonable to hypothesize that persons rated high on the variable adapt their goals to the expectations of other people, whereas those low on the variable, the more "inner directed," persist in goals set for themselves.

As previously pointed out the factorial results have proven a useful method for organizing the data. Composite scores for each factor were developed by selecting variables with the higher loadings on each factor (generally .30 or higher) and simply adding the variable scores. The resulting values are not "factor scores" because no attempt was made to partial out other factors, particularly the general ones, which contribute to each score. It was decided, however, that the composite scores thus derived would tend to reflect the underlying factors. Applications of these scores are discussed in the next section of this report.

ANALYSES OF TECHNIQUES

Several studies have been made of the techniques used in collecting information on the personal characteristics of the participants. Particular attention has been given to the contributions the methods have made to the evaluations by the assessment staff. Not all of the methods used have been studied, though eventually it is planned to make the coverage as complete as possible. The methods studied to date include the more directly scorable; that is, the group exercises, In-Basket, mental ability tests, and personality questionnaires. Major omissions are the less easily quantified interviews and projective instruments.

Group Exercises

Each of the group exercises (Manufacturing Problem and Group Discussion) was observed by two members of the assessment staff. These observers made notes from which a report was prepared on each participant for presentation to the assessment staff. The observers also independently rated and ranked each participant on his overall contribution to the problem. In addition, for the Group Discussion only, the observers rated each man on his effectiveness in oral presentation.

Additional ratings and rankings were obtained from each participant who evaluated his own performance (self-rating and -ranking) and the performance of each man in the group (peer rating and ranking) on his overall contribution to the problem. For analysis purposes the peer evaluations were averaged. Furthermore, as indications of "self-objectivity," difference scores between each self-rating and -ranking and the corresponding average peer rating and ranking were ascertained.

In summary, the following scores were obtained on each participant and presented, along with a descriptive report, to the assessment staff at its evaluation meeting:

- Manufacturing Problem
  - Ratings on overall contribution
    - Observers (independent and average)
    - Peers (average)
    - Self
    - Algebraic differences, self and average peer
  - Rankings on overall contribution:
    - Observers (independent and average)
    - Peers (average)
    - Self
    - Algebraic difference, self and average peer
- Group Discussion
  - Ratings on oral presentation:
    - Observers (independent and average)

Ratings on overall contribution:
- Observers (independent and average)
- Peers (average)
- Self
- Algebraic difference, self and average peer

Rankings on overall contribution:
- Observers (independent and average)
- Peers (average)
- Self
- Algebraic difference, self and average peer

The various scores resulting from the two group problems were analyzed by correlational methods in order to ascertain:

1. The extent of agreement between raters (observers, peers, and self).
2. The extent of overlap between methods of evaluation (ratings and rankings).
3. The extent of overlap between evaluations of performance in the two exercises.
4. The extent to which the exercises contributed to the evaluations of the men assessed (ratings by the assessment staff).

In determining rater agreement, overlap between methods, and overlap between exercises the results for 355 participants in five companies were combined. Product-moment correlations between the variables were computed for each company sample. The correlations were then averaged (after converting to z's) in order to obtain estimates for the entire sample.

Rater Agreement

The extent of agreement between raters is shown in Table 1. The evaluations (ratings and rankings) of Observer 1 were correlated with those of Observer 2. In addition, the averaged ratings and rankings of the two observers were correlated with the averaged ratings and rankings of the peers and the self-ratings and -rankings. Finally, the averaged peer ratings and rankings were correlated with the self-evaluations.

TABLE 1

RATER AGREEMENT

	Observers (1 with 2)	Observers with Peers	Observers with Self	Peers with Self
Manufacturing Problem				
Overall rating	.60	.64	.47	.45
Overall ranking	.69	.59	.43	.38
Group Discussion				
Overall rating	.75	.73	.55	.51
Overall ranking	.75	.69	.50	.45

For both exercises the correlations are positive and relatively high. The agreement between raters tends to be higher, in general, for the Group Discussion than for the Manufacturing Problem. The self-ratings tend to correlate lower with the peer and observer ratings than do the observers with each other or with the peers.

The need for multiple observers is indicated by the magnitude of agreement between the two observers. Though relatively high, it is not sufficiently high to warrant dispensing with either of the observers.

The reasonably high agreement between the different sources of ratings suggests that all were reacting to many of the same aspects of individual performance. Evidently, the objective aspects of both exercises are sufficiently apparent to observers and participants alike to influence them similarly in arriving at their evaluations.

Method Agreement

The extent of agreement between the rating and ranking methods of evaluation is presented in Table 2. For both the observers and peers the average ratings were correlated with the average rankings.

TABLE 2

METHOD AGREEMENT

	Observers	Peers	Self
Manufacturing Problem	.89	.84	.64
Group Discussion	.89	.83	.72

As might be expected, the correlations are relatively high. The ratings and rankings yield quite similar results. Here again, however, the correlations for self-judgments are lower.

Overlap of Exercises

Table 3 shows the relationships between the evaluations of performance in the two exercises.

TABLE 3

Correlations between Manufacturing Problem and Group Discussion

	Observers	Peers	Self
Overall rating	.45	.52	.38
Overall ranking	.41	.46	.38

The relationships shown are positive and fairly high. The two exercises elicited some common aspects of behavior in spite of the different nature of the two techniques—one a small-business game and the other a group discussion. Yet the size of the correlation also indicates that each exercise makes a unique contribution.

Oral Communications Skills

As noted previously, an additional evaluation was obtained from the Group Discussion. The observers rated each participant on his performance in making an oral presentation. The relationships of these ratings to selected variables are presented in Table 4. The ratings were averaged prior to computing the correlations with other variables.

It will be noted that the agreement between observers (.54) is markedly lower than the .75 shown in Table 1 for total contribution to the exercise. Performance over an hour-long period of group interaction appears easier to judge than a short talk. The lower reliability of these ratings undoubtedly reduces the correlation with overall performance in the group discussion, but, even so, it is apparent that skill in oral presentation is only one factor in effectiveness in the discussion problem.

TABLE 4

Oral Communications Skills
Correlations of Observer Ratings

	r
Observer 1 with Observer 2	.54
Group Discussion	
Observer ratings (overall)	.54
Observer rankings (overall)	.52
Peer ratings	.43
Self-ratings	.29
Manufacturing Problem	
Observer ratings	.31

Contributions to the Staff Evaluations

In order to assess the contributions of the group exercises to the total assessment process the ratings of performance in each exercise were correlated with the final judgments of the assessment staffs. These judgments are reflected in the scores based on the factors previously described and in the prediction of advancement potential. The factors, as will be recalled, are as follows:

Factor	Identification
I	General effectiveness
II (college graduates only)	General effectiveness
III (college graduates only)	Passive dependency
IV	Administrative skills
V	Interpersonal skills
VI	Control of feelings
VII	Intellectual ability
VIII	Work-oriented motivation
IX	Passivity
X	Dependency
XI (college graduates only)	Nonconformity

The correlations of the various ratings for the group exercises and the staff judgments are shown in Table 5. The college and noncollege samples are treated separately.

In general, with a few exceptions, the observer ratings correlate highest with the factor scores derived from the staff ratings. The average peer ratings are next highest while the self-rating have the lowest correlations. The self-rating correlations are notably low in many instances.

For the college graduate sample the correlations of the staff ratings with observer ratings in the group exercises are noticeably higher for the Group Discussion than for the Manufacturing Problem. The staff seems to have "gotten more" out of the discussion. For the noncollege sample the very small differences that exist favor the Manufacturing Problem.

The correlations of the ratings and rankings from the two exercises are nearly always higher with the general evaluations of the candidates—Factors I and II and staff prediction—than with the more spe-

TABLE 5

Correlations of Groups Exercises with Staff Judgments

	Factor											Staff prediction
	I	II	III	IV	V	VI	VII	VIII	IX	X	XI	
College sample ($N = 207$)												
Manufacturing Problem												
Observer rating	44	42	−29	31	39	37	18	30	−35	−18	19	41
Peer rating	33	31	−24	24	29	22	11	28	−28	−15	09	39
Self-rating	15	07	−23	03	11	13	04	10	−26	−17	10	11
Group Discussion												
Observer rating (overall)	67	67	−38	48	62	47	27	45	−39	−22	24	60
Observer rating (oral presentation)	56	60	−24	42	52	32	26	34	−29	−07	17	49
Peer rating	58	57	−35	41	52	36	31	40	−37	−20	31	52
Self-rating	29	26	−24	19	23	21	10	18	−28	−14	20	18
Noncollege sample ($N = 148$)												
Manufacturing Problem												
Observer rating	60			51	52	35	20	39	−34	−17		42
Peer rating	51			48	47	27	22	20	−31	−16		40
Self-rating	38			37	23	13	10	19	−43	−08		31
Group Discussion												
Observer rating (overall)	57			41	45	36	15	36	−36	−21		38
Observer rating (oral presentation)	56			42	47	20	12	36	−41	00		47
Peer rating	56			44	47	24	17	33	−38	−17		42
Self-rating	42			33	27	31	11	22	−30	−33		34

cific factors. This is perhaps not surprising since what is rated in the group problems is general effectiveness.

Because the two exercises are group-interaction exercises it might be expected that among the more specific factors the "interpersonal skills" factor scores would show the highest correlations with the observers' performance ratings. This proved to be the case, but there were almost equally large correlations with other factors, including Administrative skills, Control of feelings, Work-oriented motivation, and Passivity. The group exercises apparently yielded evidence on multiple aspects of performance.

In-Basket

Because the In-Basket exercise was not "scored" in any way during assessment a method for quantitatively evaluating the reports for research purposes was evolved. This step was essential for determining relationships between the results of the exercise and the later judgments of the assessment staff.

The method used consisted simply of asking two members of the research staff to independently read each narrative report which had been prepared at the assessment center and rate overall performance on a 5-point scale. A score of 3 indicated an average performance, 1 and 2 below-average performance, while 4 and 5 signified above-average performance. Following the independent ratings a composite rating for each man, resolving discrepancies, was reached by mutual agreement.

To obtain an estimate of rater agreement the correlation between the independent evaluations of the raters was determined. An r of .92 resulted, indicating a high degree of agreement. Both raters stated that with a few exceptions the reports clearly evaluated the performances of the men assessed, so that there was little difficulty in assigning the ratings.

The results of correlating the ratings of the In-Basket reports with the assessment staff evaluations are shown in Table 6.

The In-Basket is primarily an admin-

TABLE 6

CORRELATIONS OF IN-BASKET WITH STAFF JUDGMENTS

	College graduates ($N = 207$)	Noncollege ($N = 148$)
	r	r
Staff prediction	.55	.51
General effectiveness (I)	.60	.59
Administrative skills	.76	.68
Interpersonal skills	.45	.49
Intellectual ability	.36	.27
Control of feelings	.39	.24
Work-oriented motivation	.44	.26
Passivity	−.18	−.27
Dependency	−.15	.00
Nonconformity	.17	—

istrative exercise, and the table shows that the technique has its highest correlations with staff judgments of administrative skills. In contrast to the results for the group exercises, these correlations are higher than those with the more general factors. This suggests that the In-Basket is a somewhat more "focused" technique than the group problems. Nevertheless, there are also substantial correlations with several factors other than administrative skills.

Mental Ability Tests

The three mental ability tests used in assessing participants were employed very specifically by the staff in making its judgments. Judgments of "scholastic aptitude" were based on scores on the School and College Ability Test (SCAT) and Critical Thinking in the Social Science Test. Scores on the Contemporary Affairs Test were taken into account in evaluating "range of interests."

Correlations of the scores on these tests with the staff evaluations appear in Table 7. The results for the college graduate (C) and the noncollege (NC) samples are shown separately.

It is hardly surprising to note that the correlations of the three tests with Intellectual Ability are generally high. For the noncollege sample the correlations of SCAT and Critical Thinking with Administrative skills also are high, probably because the variable "scholastic aptitude" was included in the scoring of this factor, though was not so included for the college graduate sample.

The correlations of the tests with the Staff prediction and General effectiveness are relatively high with the SCAT Verbal tending to have the highest correlations of the various scores across samples. Though the mental ability tests make important contributions to the staff judgments it is apparent, however, that there is much variance in the judgments that cannot be explained by scores on these tests.

TABLE 7

CORRELATIONS OF MENTAL ABILITY TEST SCORES WITH STAFF JUDGMENTS

	School and College Ability Test						Contemporary affairs		Critical thinking in social science	
	Verbal		Quantitative		Total					
	C	NC	C	NC	C	NC	C	NC	C	NC
Staff prediction	36	44	06	29	27	41	29	22	31	43
General effectiveness (I)	47	50	14	39	39	50	40	27	39	51
Administrative skills	37	69	16	58	34	72	30	36	34	68
Interpersonal skills	22	23	−03	15	13	22	19	12	20	28
Intellectual ability	79	64	29	46	70	62	73	55	65	54
Control of feelings	23	32	09	20	21	28	20	23	16	27
Work-oriented motivation	17	14	05	22	14	22	14	−02	14	17
Passivity	−16	−21	00	−17	−10	−21	−10	−07	−16	−22
Dependency	−17	−16	−20	−15	−24	−15	−14	−20	−19	−08
Nonconformity	48	—	13	—	40	—	53	—	48	—

Note.—C = college graduate: NC = noncollege.

Personality and Attitude Questionnaires

With one exception scores on the personality and attitude questionnaires were used much more generally than were the mental ability test scores in judging the participants. The scores (24 in all) were read at the staff evaluation meetings and each staff member was expected to draw his own inferences from the results. The exception was the Authoritarianism score from the Opinion Questionnaire which specifically influenced judgments of "social objectivity."

Correlations of the questionnaire scores with the staff judgments are shown in Table 8 for the college graduate and noncollege samples. Only the higher correlations (.20 or greater for either sample) are shown. The questionnaires are coded as follows:

PPS—Edwards Personal Preference Schedule
GAMIN—Guilford-Martin Inventory of Factors GAMIN
OQ—Opinion Questionnaire
SATL—Survey of Attitudes Toward Life

In general, the correlations of the personality and attitude measures are distinctly lower than are those for the mental ability tests. As would be expected the questionnaire scores have their highest correlations with the motivational variables. The scores which correlate .20 or higher most often are the Edwards dominance and abasement scales and the Guilford-Martin general activity and ascendancy scales.

Variations between the two samples in the magnitudes of the correlations will be noted. These probably reflect population differences as well as differences in the rating variables "scored" for each factor. For that matter the patterns of correlations appear to "fit" the factors for the college sample better than they do for the noncollege sample. This is particularly true of the dependency factor.

The relatively low correlations of the personality and attitude questionnaire variables suggest that the assessment staffs may have been more influenced in their judgments of motivational characteristics by data obtained from the interview and projective instruments than by question-

TABLE 8

CORRELATIONS OF PERSONALITY AND ATTITUDE QUESTIONNAIRE SCORES WITH STAFF JUDGMENTS

	College graduates r	Non-college r
Staff prediction with		
PPS dominance	.29	.24
GAMIN general activity	.20	[a]
PPS abasement	−.15	−.20
General effectiveness (I) with		
PPS dominance	.33	.22
GAMIN general activity	.28	[a]
GAMIN ascendance-submission	.27	[a]
PPS abasement	−.20	−.22
PPS nurturance	−.08	−.21
PPS deference	.01	−.20
Administrative skills with		
PPS dominance	.30	.30
PPS abasement	−.12	−.24
PPS nurturance	−.03	−.22
Interpersonal skills with		
PPS dominance	.27	.14
PPS abasement	−.07	−.20
Intellectual ability with		
PPS abasement	−.23	−.11
Control of feelings with		
GAMIN ascendance-submission	.28	[a]
PPS dominance	.26	.13
PPS succorance	−.23	−.12
SATL total	.20	.08
PPS exhibition	.06	.25
PPS abasement	−.09	−.20
Work-oriented motivation with		
PPS dominance	.28	.03
PPS deference	.22	−.17
PPS achievement	.21	.18
Passivity with		
GAMIN general activity	−.43	[a]
GAMIN ascendance-submission	−.35	[a]
PPS dominance	−.27	−.29
SATL total	−.27	−.22
PPS abasement	.20	.23
PPS intraception	−.02	−.22
PPS deference	.04	.20
Dependency with		
PPS achievement	−.29	.04
PPS succorance	.24	.04
GAMIN general activity	−.23	[a]
GAMIN ascendance-submission	−.23	[a]
PPS dominance	−.22	−.04
PPS abasement	.21	.05
PPS nurturance	.21	.11
SATL total	−.21	.11
GAMIN inferiority feelings	−.20	[a]
PPS aggression	−.20	−.17
PPS deference	.12	.22
OQ authoritarianism	.00	−.34
OQ negativism	−.06	−.23
Nonconformity with		
OQ authoritarianism	−.41	
OQ acquiesence	−.29	
PPS succorance	−.29	
PPS change	.22	
PPS abasement	−.21	
PPS order	.20	

[a] Not administered to entire sample.

naire results. Further studies will be necessary to ascertain whether this is so. In addition, the reliabilities of some of the questionnaire variables may be relatively low, which could result in reduced estimates of the "true" correlations of the underlying variables.

Relative Contributions

To obtain perspective on the relative contributions made by the various assessment methods the correlations of each method with the staff judgments were examined and the highest correlations selected. These correlations are shown in Table 9.

The purpose for selecting the highest correlations was to obtain indications of the maximum contributions made by each method to the evaluations. These correlations, of course, may be underestimates of the total variance accounted for by any one method but are reasonably indicative of the relative contributions. For each of the rating variables the method correlating highest with it is italicized.

The comparisons show that some of the methods contributed more than others to the staff evaluations. The simulations—group problems and In-Basket—show generally higher correlations than the paper-and-pencil devices. Among the latter, the mental ability test shows up, on the average, stronger than the personality questionnaire. All the techniques, however, show a good correlation with at least one factor.

The table shows also that the five techniques account for more of the variance in some factors such as Administrative skills (IV), Interpersonal skills (V), and Intellectual ability (VII) than in others like Control of feelings (VI), Work-oriented motivation (VIII), Passivity (IX), or Dependency (X). It may be that some of the latter evaluations depend heavily on the interview or the projective tests.

In order to ascertain the relative independent contributions of the more highly correlating methods to the overall judgment of the assessment staffs the methods selected were intercorrelated and multiple-correlation coefficients and regression weights against staff predictions for the two samples were computed.

The four methods selected, along with the essential statistical information, are shown in Table 10. The correlations shown for the two group exercises are based on the observer ratings. SCAT 1A Verbal was selected as the measure of mental ability because its correlations with the staff prediction were the highest of such measures.

The four methods combined account for 56% of the variance of the staff predictions in the college sample and 44% in the noncollege sample. The regression weights vary, the Group Discussion and In-Basket having

TABLE 9

HIGHEST CORRELATION EACH ASSESSMENT METHOD WITH STAFF JUDGMENTS

	Staff prediction	Factor								
		I	IV	V	VI	VII	VIII	IX	X	XI
College graduates										
Manufacturing Problem	41	44	31	39	37	18	30	−35	−18	19
Group Discussion	*60*	*67*	48	*62*	*47*	31	*45*	−39	−22	31
In-Basket	55	60	*76*	45	39	36	44	−18	−15	17
Mental ability test	36	47	37	22	23	*79*	17	−16	−24	*58*
Personality questionnaire	29	33	30	27	28	−23	28	*−43*	*−29*	−41
Noncollege										
Manufacturing Problem	42	*60*	51	*52*	35	22	*39*	*−43*	−17	
Group Discussion	47	57	44	47	*36*	17	36	−41	−33	
In-Basket	*51*	59	68	49	24	27	26	−27	00	
Mental ability test	44	51	*72*	28	32	*64*	22	−22	−20	
Personality questionnaire	24	22	30	−20	25	−19	18	−29	*−34*	

TABLE 10

INTERCORRELATIONS AND REGRESSION COEFFICIENTS

	1	2	3	4	5	Beta
College graduates						
1. In-Basket	—	.17	.29	.26	.55	.37
2. Manufacturing Problem	.17	—	.40	.16	.41	.17
3. Group Discussion	.29	.40	—	.20	.60	.38
4. SCAT 1A verbal	.26	.16	.20	—	.36	.16
5. Staff prediction	.55	.41	.60	.36	(R = .75)	
Noncollege						
1. In-Basket	—	.28	.25	.30	.51	.33
2. Manufacturing Problem	.28	—	.46	.22	.42	.19
3. Group Discussion	.25	.46	—	.16	.38	.17
4. SCAT 1A verbal	.30	.22	.16	—	.44	.27
5. Staff prediction	.51	.42	.38	.44	(R = .66)	

the greatest weights in the college sample and the In-Basket and SCAT Verbal in the noncollege sample. Each of the methods makes a unique contribution, however, to the predictions of the assessment staffs.

The analysis also makes it clear that the three situational evercises had a major influence on the judgments of the assessment staffs. The three combined account for 50% of the variance in the staff predictions for the college sample and 31% for the noncollege sample. In contrast the mental ability measure, SCAT Verbal, accounts for only 6% and 12%, respectively, of the variance.

PREDICTION OF PROGRESS

Because the Management Progress Study has been in existence for only 9 years and because the participants in the Study were assessed over a 4-year span (1956–1960), it would be presumptuous to expect maximally discriminating criteria of progress in the management hierarchy to have become available yet. Furthermore, the assessment procedures used in the first telephone company to participate in the Study (summer of 1956) were sufficiently revised so that the assessment data obtained in that company are of little value for comparison purposes. Consequently, progress data spanning 8 years or less were obtainable for the analyses to be described.

In July 1965, five of the participating telephone companies submitted information on the progress made up to that date by the men in the Study. The data included the management level achieved and current salary. Table 11 summarizes these data,

TABLE 11

PROGRESS IN MANAGEMENT

Sample	N	Year assessed	Educational background	Percentage at each management level (6/30/65)		
				3–4	2	1
A	54	1957	College	43	55	2
B	83	1958	Noncollege	7	33	58
C_1	27	1959	College	15	78	7
C_2	39	1959	Noncollege	18	67	15
D_1	19	1960	College	32	63	5
D_2	22	1960	Noncollege	23	36	41
E	25	1960	College	16	68	16
Combined	125	1957–60	College	30	64	6
Combined	144	1957–60	Noncollege	13	42	45
Combined	269	1957–60	College and noncollege	21	52	27

showing the sample (coded), educational backgrounds of the participants, the year in which the men were assessed, and the numbers at each level.

Levels 3 and 4 are the "middle-management" levels in the Bell System. Level 3 is the objective level for which those college graduates who are classified as management trainees were employed. They were expected to achieve this level within a reasonable period of time (5 to 10 years).

Approximately one-fifth (21%) of the men assessed who were still employed in 1965 had achieved middle-management status. Variations between samples are marked, ranging from a high of 43% in A to a low of 7% in B. A slight majority (52%) of those assessed have achieved the second level of management, whereas a fourth (27%) are still at the first level. Again, variations between samples can be noted, ranging from 2% at the first level in A to 60% in B.

The college graduates generally have progressed more rapidly than the noncollege men. This is not surprising since all but one or two of the college men were employed as having middle management potential while considerably more of the noncollege men were not so appraised by line management at the time of assessment. Whereas 30% of the college graduates have achieved middle management only 13% of the noncollege men have done so. Conversely, 45% of the noncollege men are still at the first level of management while only 6% of the college graduates have failed to achieve a higher level.

For each of the samples relationships between management level obtained and assessment staff predictions (will achieve middle-management in 10 or less years)

TABLE 12

RELATIONSHIPS OF STAFF PREDICTIONS TO PROGRESS

Sample	Staff prediction (will make middle management)		Management level (1965) 3-4 %	2 %	1 %	Significance (P)
A	Yes	33	58	39	3	.02
	No or ?	21	19	81	0	
B	Yes	20	30	70	0	.001
	No or ?	63	0	21	79	
C_1	Yes	11	27	73	0	.17
	No or ?	16	6	81	13	
C_2	Yes	13	38	62	0	.03
	No or ?	26	8	69	23	
D_1	Yes	10	50	50	0	.09
	No or ?	9	11	78	11	
D	Yes	8	24	38	38	—
	No or ?	14	21	36	43	
E	Yes	8	37	63	0	.08
	No or ?	17	6	71	23	
Combined college	Yes	62	48	50	2	.001
	No or ?	63	11	78	11	
Combined noncollege	Yes	41	32	61	7	.001
	No or ?	103	5	35	60	
All samples combined	Yes	103	42	54	4	.001
	No or ?	166	7	51	42	

are shown in Table 12. For analysis purposes the data were dichotomized (grouping first and second levels). Appropriate significance tests were then applied to determine whether the observed relationships are statistically reliable. The chi-square test was used with Samples A and B and the combined samples while, with one exception, Fisher's exact test (Siegel, 1956) was applied to the remaining samples. For the exception, Sample D2, no statistical test was applied, because the relationship observed is inconsequential.

Although the results vary from sample to sample, the most marked relationships being obtained for the two samples having the longest service in management since being assessed, they show that the assessment staffs were clearly able to identify those more likely to advance in their organizations. Of the 55 men achieving middle management, 43 (78%) were predicted correctly by the assessors. In contrast, of the 73 men who have not advanced beyond the first level of management the assessment staffs predicted that 69 (95%) would not reach middle management within 10 years.

Insufficient time has elapsed for completely evaluating the predictive accuracy of the assessment staffs. Because the predictions are for 10-year periods (to achieve middle management), it will not be until 1970 that an evaluation can be made on all of the men in the Study for the specified time period. By that year all of the men assessed will have completed at least 10 years of service in management since being assessed.

Specific Variables

In interpreting the factorial analysis of the assessment variables it was noted that judgments of general effectiveness appeared to account for much of the variance in the staff ratings. It was also pointed out that the staffs were able to make many discriminations on more specific variables. It is of interest to determine, therefore, whether general impressions or judgments on more specific variables are the more predictive of progress in management.

The criterion measure used in making this analysis is salary progress (determined by taking the difference between salary on June 30, 1965, and salary at time assessed). This measure has the advantage for correlation purposes of being more discriminating than management level. Furthermore, it is not as dependent as is current salary on previous salaries (in this instance on salary at time assessed). In the seven samples studied the correlations between management level and salary progress range from .38 to .84 with a median r of .71, indicating that despite the restricted range of levels the overlap between current level and salary progress is substantial.

The correlations between the derived "factor" scores (based on the assessment variables) and salary progress appear in Table 13. In addition, correlations for the situational exercises, ability tests, and personality questionnaires are shown.

Because the correlations for three of the samples (D1, D2, and E) seem low and erratic, interpretation of these data focuses on samples A through C2. The men in the latter samples have had at least 6 years of service in management since being assessed. Presumably the measure of salary progress for these samples is sufficiently stable to yield meaningful correlations with the various predictions used.

Though the correlations of each of the judgment variables vary considerably across the four samples certain consistencies can be noted. For one, the overall ratings of the staffs on general effectiveness do indeed have the highest correlations with salary progress (median r of .48 compared to the next highest median of .39). Secondly, some more specific characteristics appear more important than others in predicting success in management. Thus, administrative and interpersonal skills, intellectual ability, lack of passivity and control of feelings appear to be more highly correlated with progress in management than do the other variables, particularly dependency which has relatively low correlations. Firmer conclusions on the relative importance of individual characteristics can be drawn, however, when a more "optimal" criterion of progress in management becomes available.

TABLE 13
CORRELATIONS WITH SALARY PROGRESS

Predictor variable	Sample						
	A (N = 54)	B (N = 83)	C_1 (N = 27)	C_2 (N=39)	D_1 (N=19)	D_2 (N=22)	E (N=25)
Staff judgment							
General effectiveness (I)	41*	45*	51*	52*	24	13	34
Administrative skills	33*	57*	24	45*	32	−11	24
Interpersonal skills	26	34*	36	33*	29	28	40*
Control of feelings	34*	17	50*	32*	20	04	00
Intellectual ability	48*	31*	30	07	30	−13	18
Work-oriented motivation	16	29*	20	41*	05	35	15
Passivity	−30*	−41*	−33	−41*	21	−15	−40*
Dependency	−25	−01	−25	01	07	24	07
Nonconformity	34*	—	32	—	16	—	20
Situational exercises							
Manufacturing Problem	15	37*	41*	50*	14	29	−01
Group Discussion	30*	33*	50*	28	26	10	38
In-Basket	27*	44*	−01	22	03	−19	28
Ability Test							
SCAT verbal	36*	35*	51*	30	19	−44*	14
SCAT quantitative	23	44*	−04	19	09	−10	−28
SCAT total	38*	45*	32	28	18	−30	−03
Critical thinking in social science	26	46*	−21	36*	−02	−38	29
Contemporary affairs	35*	26*	32	−09	32	−17	37
Questionnaire							
Edwards PPS							
ach	20	09	12	25	−15	−28	−10
def	−03	09	13	−19	−06	02	42*
ord	−05	−01	−09	21	15	39	−23
exh	−03	−03	20	25	−38	−21	17
aut	−09	−01	−04	−07	01	−55*	04
aff	−12	−11	−01	00	−25	16	18
int	14	24*	05	14	02	35	12
suc	−14	−29*	−15	−16	08	09	−19
dom	26	40*	01	−05	10	19	27
aba	−32*	−25*	13	11	−08	−11	07
nur	−02	−28*	02	−14	−23	30	−01
chg	01	10	−41*	−11	13	−22	03
end	02	−07	−39*	03	15	39	00
het	01	−05	13	−15	13	−01	−39
agg	15	17	49*	01	13	−43	−03
Guilford-Martin							
G	−03	—	24	02	32	25	22
A	20	—	26	12	35	22	09
M	17	—	−45*	−05	14	13	06
I	19	—	−03	08	−10	16	−03
N	−07	—	−02	18	−36	22	05
Attitudes toward life	08	05	52*	−06	08	23	−01
Opinion questionnaire							
A	02	−24*	10	−15	−22	12	−25
a	−18	−02	−07	−27	−19	31	−14
n	08	−06	05	09	−01	−25	−42*

* P less than .05 that r = .00.

Assessment Techniques

That the various assessment methods also vary in predictive accuracy is apparent from inspecting the data in Table 13. The situational exercises and the paper-and-pencil ability tests have higher correlations across the four samples than do the personality questionnaires.

Among the situational exercises the Manufacturing Problem has the higher correlations while the In-Basket has the lower ones. The SCAT, particularly the Verbal part, has the higher correlations among the ability tests.

Because of the cost of assessment procedures a question could be raised regarding the gain obtained from using such procedures over the use of much simpler ones, for example, paper-and-pencil ability tests. Though the data in Table 13 do not provide a definitive answer to such questions they offer clues.

The correlations of the assessment ratings, particularly the overall ones on general effectiveness, do tend to be higher than are the correlations for the more specific variables or for individual techniques. Secondly, the correlations for the more elaborate situational exercises compare favorably to those of the ability tests. Finally, when mental ability, measured by a paper-and-pencil test, is partialled out of judged ability reliable variance remains.

The results of the latter analysis are shown in Table 14. For each sample the staff judgment variable and the ability test correlating highest with salary progress were selected. Partial correlations then were computed with the ability test being partialled out of the correlation between the staff judgment and salary progress. The t statistic was used to test for the reliability of the partial correlations.

In three of the four samples, reliable variance remains after partialling out the test scores. The results thus indicate that the assessment process does contribute more than can be gained by the simple administration of paper-and-pencil ability measures.

Discussion

Though much research on the assessment process in the Management Progress Study remains to be done, the results so far obtained are informative regarding:

1. The nature of the assessment staff evaluations.

2. Contributions to staff judgments of the techniques employed.

3. The "validity" of the evaluations.

Previously reported research on the use of assessment methods has focused on the "predictive validity" of such methods. Relatively little information has been generated on the assessment process per se.

In this regard recognition should be given the authors of the OSS (1948) report who place considerable emphasis on the entire assessment process. Much of their discussion, however, is either theoretical or descriptive. Relatively little data regarding the nature of the judgments made or the contributions of the various techniques employed are presented.

Many of the published studies do offer considerable information regarding the predictive validities of the various techniques used in assessing their subjects. For that matter, despite contrary theoretical considerations, several investigators have treated the techniques on a par with the

TABLE 14

PARTIAL CORRELATIONS WITH SALARY PROGRESS

	A		B		C_1		C_2	
	r	Partial r	r	Partial r	r	Partial r	r	Partial r
Staff judgment	.48	.32*	.57	.39**	.51	.29	.52	.42**
Ability test	.38		.46		.51		.36	

* p less than .05 that $r = .00$.
** p less than .01 that $r = .00$.

judgments of the assessment staffs. The study of Air Force officers (MacKinnon et al., 1958) can be cited as an example. In the report of this study the correlations between over 600 predictors and several criteria are shown. Of the predictors, only a relative few reflect the judgments of the assessment staff.

Nature of the Judgments

From the factorial analyses of the ratings by the Management Progress Study assessment staffs it is apparent that the staffs were influenced considerably by their overall judgments of the men assessed, particularly in evaluating potential for advancement in management. On the other hand, the staffs did make many intraindividual discriminations; the ratings reflect much more than "halo."

"Halo" in rating may reflect the inability of the raters to make discriminations among the various characteristics evaluated. The factorial results from the Study data, indicate, however that the judgments of the assessors were based on observed consistencies in the behaviors of the men assessed and on their judgment regarding the relative importance for managerial potential of the various characteristics rated. This inference is supported by the intercorrelations between evaluated performances in different techniques (see Table 10), by the varied loadings on the general factors (Appendix B), and by the fact that for each sample much of the variance in the staff ratings of potential is accounted for by the general factor.

The fact that the assessment staffs made many discriminations among the characteristics rated is further evidence that the "halo" inferred from the factorial results is a resultant of other than rater error. Actually, the factorial results do not completely reflect the extent of the discriminations made. Once the reliabilities of the rating variables are determined it will be possible to ascertain the uniqueness of each variable.

Comparisons of the factorial findings for the Management Progress Study with the results of published studies are limited because relatively few similar analyses have been reported, because methods in making such analyses have varied, and because variations in the numbers and kinds of variables on which the assessment staffs made judgments. In the OSS (1948) study four factors were obtained from 11 rating variables. Kelly and Fiske (1951) report nine first-order factors and five second-order from analysis of 42 variables. In the study of Air Force officers (MacKinnon et al., 1958) a cluster analysis of 30 variables yielded three clusters. Holt and Luborsky (1958) report three factors from each of two separate analyses of 20 variables.

Furthermore, few attempts have been made to ascertain the extent of "halo" in the ratings. In the Holt and Luborsky (1958) analyses general factors do account for much of the variance obtained in each analysis. A median correlation of .42 between the factors (oblique rotation) reported by the OSS (1948) suggests that much of the variance could have been accounted for by a general factor.

Some effort has been made to ascertain the reliabilities of assessment staff judgments. Kelly and Fiske (1951), in particular, report much evidence regarding such reliabilities. Though estimates of the reliabilities of individual assessors in the Management Progress Study are yet to be determined, the magnitudes of the communalities obtained from the factor analyses indicate that the pooled ratings for many of the variables are reasonably reliable.

Contributions of the Techniques

The data reported make it apparent that the situational techniques (group exercises and In-Basket) used in the Management Progress Study produced, despite their complexities, reasonably reliable results and that they markedly influenced the judgments of the assessment staffs. The paper-and-pencil instruments had less influence on staff evaluations generally, though they did influence them in many specific ways.

The findings further indicate that neither kind of technique could have been omitted without loss of important information. All of the methods so far investigated appar-

ently contributed some unique information. A better evaluation of all the techniques can be made once studies on the interview and projectives have been completed.

Economic considerations increase the importance of the findings regarding the situational techniques. These methods are costly and time consuming to administer. The data presented appear to justify the costs entailed.

Comparisons of the contributions to the staff evaluations made by the various techniques used in the Management Progress Study to results obtained by other investigators are restricted almost entirely to the OSS (1948) report. Many investigators, as has been noted, have reported the relationships of specific techniques to performance criteria, particularly supervisory ratings. Such data, however, shed little light on the assessment process.

The OSS (1948) assessors viewed the clinical interview as the nucleus of their assessment program. The interviewer prepared for the interview by reviewing a completed personal history questionnaire and the results obtained from a variety of paper-and-pencil tests. Subsequent to the interview he rated the assessee on several variables, thus facilitating eventual analysis of the data, and prepared a portion of a personality sketch of the individual for presentation at the assessment staff meeting. Consequently, it is hardly surprising that the interviewer ratings correlated highly with all of the assessment variables, being highest with 7 out of the 10 characteristics rated.

The situational techniques also were major contributors to the OSS assessment process. A "situationist" presented another portion of the personality sketch prepared on each assessee at the staff meeting, along with recommendations of the staff team which administered the situational tests. Subsequent correlational analysis resulted in relatively high correlations with the staff ratings, though ranging considerably in magnitude.

Very little data on the paper-and-pencil tests and none specifically on the projectives are given in the OSS report. Two mental ability tests correlated in the .60s with the staff rating on "Effective Intelligence."

Kelly and Fiske (1951) report a little data on relationships between paper-and-pencil scores and staff ratings. The correlations between such scores and the final rating of overall suitability are generally low. A mental ability test correlates .36 with the rating whereas scores on a variety of interest and personality questionnaires range from −.21 to .25.

From the limited amount of information available it is apparent that much more data on the contributions of various kinds of techniques to the assessment process would prove useful. For that matter, assuming the original data to be available, analyses similar to those being carried out on the Management Progress Study data would contribute to a better understanding of the assessment process.

Validity of Evaluations

The end product of an assessment process is a series of evaluations by the assessors. The success of the activity depends upon the accuracy of these judgments.

In the Management Progress Study the staff evaluations consisted of ratings on 25 characteristics deemed relevant to the purposes of the study and overall evaluations of management potential and likelihood of remaining in the employ of the Bell System. Two kinds of evidence are presented in this report which bear on the accuracy of the ratings.

The first kind can be considered "internal." It consists of factorial results and the correlations between scores for various techniques and the ratings.

The second kind can be thought of as "external." It has to do with the "predictive" validity (American Psychological Association, 1954) of the ratings. It consists of the correlations and other data showing relationships between the evaluations and subsequent progress in management.

The first kind of evidence is more suggestive than definitive. The factorial results, for example, make sense. The variables rated tend to load on the factors which one would expect them to, though the results

appear to be somewhat more reasonable for the college than for the noncollege sample.

The correlations between the various techniques and the composite ratings based on the factorial results also appear reasonable. Again, however, the findings seem to be more consistent in some instances for the college sample than for the noncollege. Whether the homogeneity of the college sample or a better "scoring" of the factors generated from this sample contributes to the apparent discrepancies has not been determined.

Some examples of the reasonableness of the correlations might be cited:

1. The relatively high correlations of the In-Basket with administrative skills, of the group exercises with Interpersonal skills, of the mental ability tests with Intellectual ability, of GAMIN General Activity (negatively) with Passivity, and of the Contemporary Affairs Test (positively) and Authoritarianism (negatively) with Nonconformity.

2. The pattern of correlations of several personality measures with Dependency (college sample, though not for noncollege sample).

More evidence, of course, would be desirable. Such evidence could come from refinements in the factor scoring and from quantifying the projectives and the interviews.

The evidence for "predictive" validity, though incomplete, is much more precise. The criterion measures used reflect progress 5 to 8 years subsequent to assessment. The criterion, though not "ultimate" (Thorndike, 1949), is of the kind referred to by Cronbach (1960) as "convergent." As such, it would be expected to correlate with a theoretically "ultimate" criterion of managerial success. In the Study, of course, "progress" in management is a criterion worthy of investigation in its own right. Presumably, many of the characteristics assessed should correlate with actual progress, as reflected by salary progress or administrative level.

The relationships between the assessment results and progress presented in this report are restricted by the relatively short period of time involved. Only a fifth of the men on whom data were obtained had achieved the third level of management ("middle management" in the Bell System). Approximately half were still at the second level. Some of the latter will advance; others may not. A more discriminating measure of progress will thus become available in a few years.

Despite the restrictions in criterion spread, however, the predictions made by the assessment staffs are quite accurate. Approximately 80% of those who have advanced to middle management were judged by the assessment staffs as having such potential. The predictions were even more accurate for those who have not advanced beyond the first level. Most of these men (95%) were judged as lacking in advancement potential. Evidently identifying the less adequate is easier than identifying those with more promise.

The results of correlating the assessment ratings, based on the factorial results, with salary progress indicates that no single characteristic determines progress in management. A composite of characteristics correlates higher across samples than do any of the more specific variables. The characteristics contributing most of the variance in the composite are administrative and interpersonal skills, intellectual ability, lack of passivity, and control of feelings. Work-oriented motivation has a lower correlation, while dependency, or lack of it, bears practically no relationship to progress.

The magnitudes of the higher correlations between the assessment ratings and salary progress compare favorably with similar correlations appearing in published reports. A few studies reporting correlations of greater magnitude have appeared (Cronbach, 1960; Trankell, 1959). Correlations of approximately the same magnitude also have been reported (Albrecht et al., 1964; Cronbach, 1960). Several investigators have reported considerably lower correlations (Campbell, Otis, Liske, & Prien, 1962; Holt & Luborsky, 1958; Kelly & Fiske, 1951; Kelly & Goldberg, 1959; MacKinnon et al., 1958; OSS, 1948).

Comparisons of this type are, of course,

limited in value. The criteria have varied as have the assessment methods used and statistical methods for evaluating the results. Furthermore, no standards exist for determining what could be considered an "acceptable" correlation. Predicting complex criterion behavior is an art that remains in the initial stages of development.

It bears repeating, also, that the progress criterion used in the Study is a developing one. It may take several years before a stable and fully discriminating criterion of progress is obtained. When achieved, refinements in the criterion can be made (e.g., adjusting for departmental differences in rates of progress) and better estimates of the predictive accuracy of the assessment ratings determined. Furthermore, assuming "error" in both the ratings and the progress criterion, analyses will be made to ascertain the locus and nature of such error.

A final note regarding the "predictive validities" of the assessment methods used should be made. The situational exercises and the paper-and-pencil ability tests are predictive of progress in management whereas none of the personality questionnaires correlate consistently with the criterion. Justification for the high cost of the assessment approach, moreover, can be obtained from the finding that the assessment ratings account for more of the variance in the progress criterion than do the simpler paper-and-pencil ability tests or, for that matter, than does any single method used.

In conclusion, it should be noted that prediction of progress in management was only one purpose of the Management Progress Study assessment centers. The other, and perhaps even more important purpose, was to provide a comprehensive picture of a fairly large number of young men as a base line for a study of the developmental changes of young adulthood.

APPENDIX A

CORRELATION MATRICES

Assessment variable	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
1 Organization and planning		74	68	52	49	41	41	45	39	36	57	37	48	53	38	44	29	06	12	41	−21	−14	−20	−05	−24	61
2 Decision making	82		67	49	47	38	47	44	42	40	52	40	44	56	33	42	33	−04	11	40	−20	−28	−26	−12	−24	59
3 Creativity	71	73		52	52	39	32	37	53	50	54	34	50	61	35	37	34	−10	24	43	−24	−21	−22	−06	−35	57
4 Human relations skills	62	65	59		69	54	49	54	24	37	49	37	57	61	40	46	33	13	20	41	−11	−12	−07	02	−30	66
5 Behavior flexibility	50	51	54	69		59	37	43	25	35	42	36	51	54	34	37	32	25	11	41	−08	−02	05	06	−23	63
6 Personal impact	41	42	53	51	50		26	27	21	19	29	28	42	44	34	40	28	21	07	34	−17	−09	05	10	−20	57
7 Tolerance of uncertainty	40	41	40	31	44	29		61	17	28	34	18	39	38	30	36	18	01	26	29	−10	−31	−26	−09	−37	39
8 Resistance to stress	44	49	49	45	50	49	54		17	20	42	34	40	45	34	41	23	13	11	38	−11	−31	−27	−22	−32	51
9 Scholastic aptitude	47	48	51	28	21	28	25	21		59	23	11	22	41	21	14	35	−12	14	27	−18	−17	−11	−13	−21	46
10 Range of interests	20	22	38	09	16	24	19	22	40		38	19	37	42	28	19	23	−11	30	37	−18	−10	−11	−01	−33	45
11 Inner work standards	36	34	36	30	28	11	19	21	27	05		59	44	45	46	52	26	20	22	31	−04	04	−09	−02	−14	46
12 Primacy of work	24	25	33	26	32	28	27	34	19	10	48		26	29	34	45	21	32	07	41	−07	11	01	−02	−07	48
13 Oral communications skills	63	59	60	61	52	56	28	38	29	24	20	32		42	34	48	29	14	16	40	−31	−14	−20	−03	−36	53
14 Perception of social cues	62	63	65	65	60	49	32	43	43	23	29	20	49		38	26	34	−01	22	34	−17	−06	−06	−02	−21	50
15 Self-objectivity	32	31	32	33	35	29	39	33	18	19	40	33	29	28		35	33	08	19	23	−05	−05	−05	−05	−19	38
16 Energy	38	39	45	39	34	40	22	30	17	16	36	48	49	29	26		25	19	04	50	−36	−13	−24	−07	−30	51
17 Realism of expectations	18	14	13	15	12	16	21	18	21	23	16	12	04	22	32	00		09	08	09	10	−01	09	01	−01	42
18 Bell system value orientation	01	−04	−01	04	16	12	20	12	02	−01	26	19	−05	09	07	02	07		−05	00	26	33	27	12	32	15
19 Social objectivity	16	10	16	13	20	13	23	17	22	12	17	09	06	21	26	01	09	26		−07	02	17	05	08	−18	04
20 Need advancement	48	51	52	37	32	33	12	31	35	16	17	32	52	31	05	40	−39	−02	05		−62	−29	−35	−30	−53	60
21 Ability to delay gratification	−33	−39	−33	−22	−06	−18	−06	−11	−23	03	05	−13	−43	−15	00	−31	36	24	04	−68		38	38	26	57	−30
22 Need for superior approval	−24	−22	−31	−04	−13	−29	−27	−39	−25	−13	08	−17	−10	−12	−12	−11	−01	13	−06	−19	25		55	43	57	−18
23 Need for peer approval	−28	−33	−33	−06	04	−14	−16	−23	−29	−17	−07	−24	−23	−02	−11	−28	04	24	10	−40	38	32		43	48	−16
24 Goal flexibility	−05	−07	−11	01	08	−06	03	−10	−06	−05	12	−12	−01	01	09	−01	−03	20	15	−09	08	24	28		24	−13
25 Need for security	−37	−39	−44	−26	−20	−25	−19	−24	−27	−27	−09	−22	−42	−31	−15	−34	17	21	−11	−47	49	27	26	12		−32
26 Staff prediction	60	62	61	55	48	56	29	42	47	28	31	31	64	50	29	31	09	06	14	53	−37	−24	−33	−10	−38	

Note.—Above diagonal—college graduates ($N = 207$): below diagonal—Noncollege ($N = 148$).

APPENDIX B

FACTOR MATRIX

Assessment variable	I		II	III	IV		V		VI		VII		VIII		IX		X		XI	h^2	
	C	NC	C	C	C	NC	C	NC	C	NC	C	NC	C	NC	C	NC	C	NC	C	C	NC
Organization and planning	53	75	50	−07	58	43	08	−05	03	−13	65	09	01	−01	−02	−09	−04	−05	−14	90	78
Decision making	55	78	48	−11	46	48	−03	−01	04	−08	−03	01	01	−03	05	−12	−09	−10	−12	79	87
Creativity	51	75	48	−07	26	24	15	05	−17	02	20	19	08	08	07	−15	−11	−13	16	70	70
Human relations skills	47	72	56	06	−05	21	38	36	15	−07	−05	−04	−04	−04	03	−02	01	11	07	72	71
Behavior flexibility	41	72	55	12	01	−02	45	38	01	14	04	−11	−01	−01	00	05	02	22	−06	69	75
Personal impact	33	62	43	08	04	−15	51	32	−16	12	−11	18	−05	−04	−03	01	−01	−14	05	61	58
Tolerance of uncertainty	45	55	34	−13	08	05	01	−11	62	48	01	−01	−14	−01	04	11	08	03	10	76	56
Resistance to stress	49	64	40	−11	−08	−05	09	07	38	38	−11	−03	14	02	06	−05	−16	−10	−06	64	58
Scholastic aptitude	33	43	23	−12	10	35	01	−11	−08	−03	67	45	−07	00	06	−02	−11	−06	−02	66	53
Range of interests	36	24	32	−06	−01	02	−01	−02	−02	−01	68	55	07	06	−06	−02	11	−14	29	80	38
Inner work standards	40	43	54	11	09	16	−01	−11	03	−11	03	03	39	60	10	19	00	19	14	66	67
Primacy of work	32	45	46	12	−09	−28	−09	−04	−03	13	−03	04	53	54	−10	−21	00	−02	−15	66	64
Oral communications skills	46	65	41	−08	08	04	20	26	10	−06	−03	14	−02	01	−15	−28	08	−05	13	49	60
Perception of social cues	45	69	49	01	21	28	12	20	16	−05	17	18	−10	−08	−06	03	13	12	−11	60	65
Self-objectivity	31	47	41	08	−06	−04	02	01	−03	18	05	13	23	29	08	14	−06	10	19	38	39
Energy	46	51	39	−10	−03	−07	04	15	12	−05	−17	−03	27	41	−29	−24	07	−12	05	58	53
Realism of expectations	19	20	34	14	−02	02	25	07	−10	10	30	32	10	09	29	60	−17	−10	00	46	54
Bell System value orientation	−05	14	27	33	−32	−13	07	−18	05	23	−11	03	28	03	06	12	01	46	−26	46	35
Social objectivity	11	22	20	07	−05	−06	−04	−12	09	22	18	21	03	01	23	−01	00	34	49	40	27
Need advancement	54	49	26	−31	−13	13	13	02	−12	−10	01	−07	17	11	−37	−69	−13	−04	−06	69	76
Ability to delay gratification	−39	−25	07	48	−13	−11	03	08	−02	07	01	05	−03	08	63	79	−03	13	−07	81	74
Need for superior approval	−36	−27	07	46	−08	−07	−06	15	−09	−36	07	−04	20	07	01	11	33	47	−26	59	47
Need for peer approval	−36	−22	07	47	−11	−09	12	19	−04	−12	12	−05	00	−06	07	27	34	60	−19	55	55
Goal flexibility	−24	−03	10	36	09	−03	14	01	03	02	08	−01	00	−04	−02	05	60	46	11	60	22
Need for security	−48	−37	−06	45	12	−18	−07	07	−15	−08	02	−20	08	−03	32	42	07	17	−36	72	43
Staff prediction	57	65	53	−07	−07	13	34	10	−04	00	22	31	03	08	−08	−33	−06	−01	−20	83	66

Note.—C—College graduates: NC—noncollege.

REFERENCES

Albrecht, P. A. Glaser, E. M., & Marks, J. Validation of a multiple-assessment procedure for managerial personnel. *Journal of Applied Psychology*, 1964, **48**, 351–359.

American Psychological Association. Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin*, 1954, **51**, No. 2, Part 2. (Supplement)

Bass, B. M. Authoritarianism or acquiescence? *Journal of Abnormal and Social Psychology*, 1955, **5**, 616–623.

Bray, D. W. The management progress study. *American Psychologist*, 1964, **19**, 419–420.

Campbell, J. T., Otis, J. L., Liske, R. E., & Prien, E. P. Assessments of higher-level personnel: II. Validity of the over-all assessment process. *Personnel Psychology*, 1962, **15**, 63–74.

Cronbach, L. J. *Essentials of psychological testing*. (2nd ed.) New York: Harper, 1960.

Holmen, M. G., Katter, R. V., Jones A. M., & Richardson, I. F. An assessment program for OCS applicants. *HumRRO Technical Report* **26**, 1956.

Holt, R. R., & Luborsky, L. *Personality patterns of psychiatrists*. New York: Basic Books, 1958.

Kelly, E. L., & Fiske, D. W. *The prediction of performance in clinical psychology*. Ann Arbor: University of Michigan Press, 1951.

Kelly, E. L., & Goldberg, L. R. Correlates of later performance and specialization in psychology. *Psychological Monographs*, 1959, **73**, (12, Whole No. 482).

MacKinnon, D. W., et al. An assessment study of Air Force officers, Parts I–V. WADC-TR-58-91 (I–V), ASTIA Documents Nos. AD 151040, AD 208700, AD 210218, AD 210219, AD 210220, Washington, D. C.: Office of Technical Services, United States Department of Commerce, 1958.

OSS Assessment Staff. *Assessment of men*. New York: Rinehart, 1948.

Siegel, S. *Nonparametric statistics for the behavioral sciences*. New York: McGraw-Hill, 1956.

Stern, G. G., Stein, M. I., & Bloom, B. S. *Methods in personality assessment*. Glencoe, Ill.: Free Press, 1956.

Taft, R. Multiple methods of personality assessment. *Psychological Bulletin*, 1959, **56**, 333–352.

Thorndike, R. L. *Personnel selection*. New York: Wiley, 1949.

Trankell, A. The psychologist as an instrument of prediction. *Journal of Applied Psychology*, 1959, **43**, 170–175.

Vernon, P. E. The validation of civil service selection board procedures. *Occupational Psychology*, 1950, **24**, 75–95.

Wherry, R. J. The control of bias in ratings. VII: A Theory of rating. *PRB Report* No. 922, Department of the Army, the Adjutant General's Office, Personnel Research Branch, 1952.

Wherry, R. J. Hierarchical factor solutions without rotation. *Psychometrika*, 1959, **24**, 45–51.

(Received April 22, 1966)

Vol. 80, No. 18 Whole No. 626, 1966

Psychological Monographs: General and Applied

MOTIVATION AND MEMORY

BERNARD WEINER[1]

University of California, Los Angeles

15 studies which examine the effects of motivation on memory are presented. It was demonstrated that the effects of motivation on retention are in part determined by the magnitude of incentive, quality of incentive, nature of the activity intervening between stimulus onset and recall, place in the memory sequence at which the motivational factor is introduced, type of stimuli, and type of experimental design. It is suggested that research in the area may require both between-Ss and within-Ss experimental designs. Rehearsal, repression, and action decrement are discussed briefly.

THE studies reported in this monograph provide evidence concerning a deceptively complex question: "Does motivation affect retention?" Two literature reviews (Rapaport, 1942; Weiner, 1966) have answered this question affirmatively. In support of his position, Rapaport cites the Lewinian studies of the recall of completed and incompleted tasks, hypnotic phenomena and hypermnesia, recall of pleasant and unpleasant experiences and the retention of stimuli associated with affective states, and the remembrance of traumatic events. Weiner also concluded that "there are studies which provide strong evidence that memory can be influenced by nonassociative factors [p. 24]." The critical demonstration experiments cited in that review pertain to the recall of high-arousal stimuli (Kleinsmith & Kaplan, 1963; Walker & Tarte, 1963); retention of stimuli associated with positive and negative incentives (Heyer & O'Kelley, 1949; Weiner & Walker, 1966); retention of affective material (Blum, 1961); and the recall of stimuli during heightened muscular tension (Bourne, 1955). (The reader is directed to Weiner, 1966, for a detailed analysis of these and related studies.)

In this paper, 15 studies are presented which examine the effects of motivation on memory. The experimental procedure in these investigations is guided by current criticisms of retention studies (Keppel, 1965; Underwood, 1954, 1964) and by a recent technique developed in the study of short-term memory (Peterson & Peterson, 1959). Underwood and Keppel have been critical of retention studies because of the confounding of learning with retention. Keppel (1965) illustrates this confusion with the following example:

> ...Peterson, Peterson, and Miller (1961) found higher recall for words than for low meaningful nonsense syllables 6 seconds following presentation. If it can be shown that differences in the recall of these items were also present on an immediate retention test, it will not be possible to determine whether the differences in the delayed retention test are to be attributed to the effect of meaningfulness on learning (estimated by the immediate retention test), or over the retention interval, or to the action of meaningfulness on both learning and retention. [p. 7]

The analysis by Underwood and Keppel lead Weiner (1966) to conclude that a number of studies in the area of motivation and memory also are subject to methodological criticism. For example, the Zeigarnik phenomenon often has been cited as supporting

[1] The author wishes to thank Robert Abramowitz, Sally Grinde, Harlan Higgins, Phyllis Kernoff, Gail Klynn, Ross Legrand, Sherrie Lindborg, Myrna Morrison, Carol Price, Robert Rosenbaum, Ellen Siegelman, and Marilyn Satuloff for their invaluable aid and many suggestions. An earlier draft of this paper was read critically by Kent Dallett, John P. Houston, and Paul Slovic. Part of this research was conducted while the author was at the Center for Personality Research, University of Minnesota. The investigation was supported in part by Public Health Service Research Grant No. MH-12603-01 from the National Institute of Health.

a hypothesized motivation-memory linkage. Yet Caron and Wallach (1957) have demonstrated that differences in the recall of completed and incompleted tasks must be attributed to differences in the degree of original learning rather than to differences in retention. A similar criticism is applicable to investigations of the relations between attitudes and retention (e.g., Levine & Murphy, 1943). A recent study by Fitzgerald and Ausubel (1963) provides evidence that differences in recall as a function of attitude toward the content of a message are produced by differential learning of that message.

In the research on memory reported here the degree of original learning between motivational and nonmotivational conditions is equated. That is, the effects of motivation on retention are disentangled from the effects of motivation on learning. Following previous suggestions (e.g., Cameron, 1947; Melton, 1963) memory is conceptualized as a multistage process. The initial period involves sensory or ideational registration and the subsequent fixation of that event. This is the stage of learning or trace formation. In the second phase of the memory process the trace of the event is latent, yet potentially available for evocation. This is often referred to as the period of trace storage. In the final process of the sequence the trace is revived by the organism. This is the stage of trace evocation or trace retrieval. Only the latter two stages, storage and retrieval, are adjudged to be memory processes. Studies of the influence of motivation on memory must be able to relate the motivational manipulation to changes occurring during either of these two stages.

A second criticism of previous studies in the area of motivation and memory is that the experiences between the occurrence of the to-be-remembered event and subsequent recall are not controlled. This also can result in a confounding of learning with retention. For example, some investigators have found that pleasant experiences are more likely to be retained than unpleasant experiences (cf. Meltzer, 1930). However, if the enhanced recall of pleasant experiences is produced by intervening rehearsal (verbal repetition), then the differential recall would be of little theoretical significance for the study of memory. Most psychologists would agree that the probability of recall of an event is a function of the number of repetitions of that event; this is a fundamental law of learning. Therefore, the behavior between the onset of a stimulus and subsequent recall must be controlled in experiments relating motivation and memory. In the investigations described in this monograph a modification of the Peterson and Peterson (1959) technique used in the study of short-term memory is employed. In that procedure the behavior of the individual between stimulus onset and stimulus recall is prescribed, and overt rehearsal is prevented.

More uncertainty is conveyed by the term "motivation" than by the word "memory." Following Atkinson (1964), this writer considers any contemporary determinant of behavior to be a motivational variable. Hull (1952) and Spence (1956) include drive, habit, and incentive among the immediate determinants of action; Lewin (1938) conceptualizes behavior to be a function of tension, psychological distance, and valence; Atkinson's (1964) model of the determinants of behavior comprises motive, expectancy, and incentive. Motivational theorists therefore conceive behavior to be a function of properties of the organism (drive level, magnitude of tension, motive strength), attributes of the environment (valence or incentive of the goal), and an associative or learning factor (habit strength, psychological distance, expectancy). Previous studies in the area of motivation and memory have related nonassociative factors pertaining to the state of the organism (arousal level, attitude, motive, etc.) and/or characteristics of the environment (message content, affective tone, magnitude and quality of incentive, etc.) to recall (cf. Weiner, 1966). In the studies described in this paper the incentive offered for retaining a stimulus is varied. Incentives were selected as the main independent variable because of the relative ease and experimental feasibility of manipulating their quantity and quality. It is believed that the general pattern of results emerging from this experimentation will be

replicable with other motivational manipulations. That is, the data are assumed to reflect general relationships between motivation and memory.

The reader must be forewarned that the purpose of this research is not merely to ascertain whether motivation influences memory. The existence of this effect was decided in an earlier paper (Weiner, 1966). It is hoped that the research will result in a specification of the conditions under which this relationship can be expected to hold. Frequently the investigator is able to isolate a relevant condition, yet the finding is not investigated in detail. It is evident that much work remains to be done after an initial demonstration. However, this program of research is directed to the exposure of a variety of factors affecting the relationship between motivation and memory, rather than to the systematic understanding of any individual determinant. To provide continuity and to convey the evolution of the research program, the experiments are reported in their historical sequence of occurrence. Often a problem appearing earlier in the program is temporarily put aside or not discovered until a later investigation.

EXPERIMENT I

The initial investigation of the series was conducted by Weiner and Walker (1966). Inasmuch as that study serves as the prototypical experiment, the experimental procedure will be repeated here in detail. In the discussion of later experiments only the modifications of that procedure will be reported.

Method

Subjects. Twenty male students enrolled at the University of Michigan participated as paid subjects.

Materials. Eighty consonant trigrams of less than 30% associative strength (Witmer, 1935) were used as stimuli. The consonants "v" and "w" were excluded; each of the remaining 18 consonants was used in no less than 10 and no more than 15 of the trigrams. The trigrams were printed on slides with one of four background colors: red, yellow, green, or white. Twenty stimuli were randomly assigned to each color.

Procedure. The subjects participated in a short-term memory task. On each trial the background color on which the trigram appeared informed the subjects of the incentive for correctly remembering the stimulus. There were four experimental conditions corresponding to the four colors: win one cent for correctly recalling the stimulus, win five cents, receive a shock for not correctly recalling the stimulus, and a control condition in which neither shock nor money was used as an incentive. The intensity of the pulse shock was 110 volts with an amperage of 60 microamps. The shock was delivered to the upper arm of the subject.

Subjects were first informed of the color-value pairings. To ensure that the value of each color was retained, the subjects were administered a 10-trial, 4-item paired-associates list consisting of the color-incentive pairs. The experimenter read the colors aloud; subjects were corrected following wrong responses. There generally were no incorrect anticipations following the third trial.

In the short-term memory task which followed, the to-be-remembered stimuli were projected on a screen for .75 second. Then slides containing random single digits were projected. There were 60 digits on each slide. The interslide interval was approximately .70 second. As an interpolated activity subjects were required to read the digits in time to a metronome which beat 3.25 times per second. There were two time intervals for the interpolated activity: 4.67 seconds and 15 seconds. Therefore, approximately 15 or 49 digits were read during the interpolated time period. Both the trigrams and the digits were read aloud. The recall time allowed between the offset of the digits and the onset of the next stimulus was 12 seconds. Recall was cued by the appearance of a blank slide on the screen. Following Trials 20, 40, and 60 there was a 10-second time delay to change slide trays.

Subjects were in all experimental conditions. In the first and last 40 trials each of the eight conditions (four incentive conditions × two time intervals) appeared five times. Within these 40 trials the order of presentation was randomized. The order of the stimuli was constant across subjects, and for all subjects the incentive value associated with a given color remained the same during the experiment. The design was counterbalanced so that every color-trigram pairing was associated with each of the incentive conditions an equal number of times. Subjects were randomly assigned to the various color-incentive combinations.

Results

In Figure 1 the percentage of correct responses is plotted as a function of incentive and the time of the interpolated activity. The analysis of variance performed on this data reveals that there is a significant main effect attributable to the incentive condition, $F(3, 57) = 6.94$, $p < .01$, and a significant interaction between the incentive

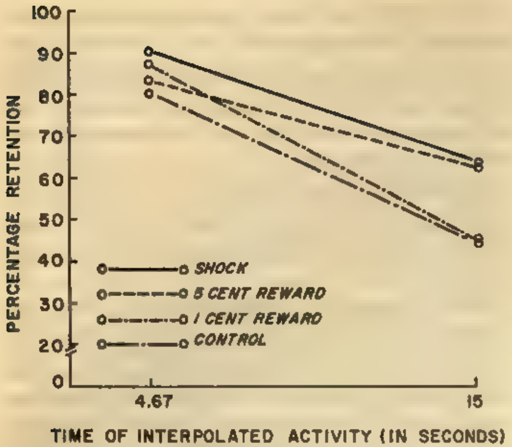

FIG. 1. Percentage retention at two time intervals as a function of incentive, employing 2 degrees of monetary reward.

condition and the time of the interpolated activity, $F(3, 57) = 4.74$, $p < .01$. Comparing retention within a time interval with a Newman-Keuls paired-mean test shows that at the shorter interval there are no significant differences between any of the means. At the 15-second time interval retention is significantly better ($p < .01$) in the five-cent and shock conditions than in either the one-cent or control conditions.

Discussion

The results support the general hypothesis that motivation affects retention. The study demonstrated that the recall of an event is in part a function of the anticipated outcome signaled by that event. Differential rates of forgetting were exhibited when the to-be-remembered stimuli were identical in the different conditions. This is important because Underwood (1964) has argued convincingly that the degree of original learning produced by intrinsic differences in the to-be-retained stimuli, for example, the number of units in the material, has been confounded with differences in retention. Further, recall did not significantly differ between conditions at the shorter time interval, and varied between 80–90%. It therefore is extremely unlikely that the differences in recall at the 15-second interval can be attributed to differences in the degree of original learning. The interaction between the time of the interpolated activity and the incentive conditions indicates that the storage of the trace is the process affected by the motivational manipulation employed in this study.

The purpose of this research is to isolate a number of factors which affect the relationship between motivation and memory. Two factors which in part influence memory have been identified in this investigation: the magnitude of an incentive (penny versus nickel), and the type of incentive (aversive and positive).

EXPERIMENT II

In the Weiner and Walker study, feedback was conveyed only when administering or withholding the shock. Conceivably the difference in recall between the shock and nonshock conditions could be attributed to differential knowledge of results (KOR). In this study further controls were instituted by providing KOR after every response. Following each response the experimenter said: "No penny" or "Penny"; "No nickel" or "Nickel"; "No shock," or shock was administered; or "Wrong" or "Right" in the appropriate condition. Other methodological changes included the substitution of purple for white as a color cue, and an extension of the short time interval from 4.7 seconds to 5.6 seconds.

Subjects were 20 male students enrolled in the introductory psychology course at the University of Minnesota.

Results

In Figure 2 the percentage of correct responses is plotted for the four incentive conditions and the two time intervals. An analysis of variance reveals that there is a significant main effect due to the experimental conditions, $F(3, 57) = 5.81$, $p < .01$. The interaction between the incentive and time interval does not reach statistical significance, $F(3, 57) = 1.44$, $p < .25$. Although this interaction is not significant, paired-mean tests are reported to allow further comparisons between the findings in Experiments I and II. A Newman-Keuls paired-mean test indicates that there are no significant differences in recall between

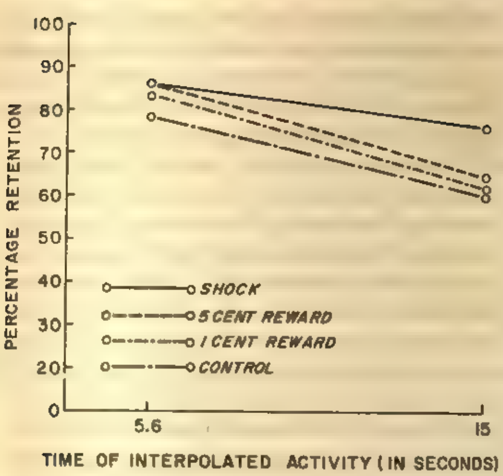

Fig. 2. Percentage retention at two time intervals as a function of incentive, employing 2 degrees of monetary reward and feedback after each response.

any of the conditions at the short time interval. At the long interval, recall of stimuli associated with shock is significantly greater ($p < .05$) than recall in the other three incentive conditions. There are no significant differences in recall at the long interval between the penny, nickel, and control conditions.

Discussion

The results partially replicate the findings in the Weiner and Walker experiment. In the present study, as in the previous investigation, the stimuli paired with shock were recalled more often than stimuli associated with a one-cent reward and stimuli for which neither shock nor money were at stake. However, this investigation failed to replicate the differences in recall between the penny-nickel and control-nickel conditions which were found previously.

The significant differences in recall between the shock and control conditions establish that differential feedback cannot account for the unequal rates of forgetting in those conditions. Further, the equality in retention at the short time interval, which varied between 78–85%, and the divergent decay rates found for identical trigrams again strongly suggest that heightened motivation during the perception of stimulus affects the subsequent availability (storage) of that stimulus.

The failure to replicate the findings in the Weiner and Walker study concerning the effectiveness of the positive reward is somewhat puzzling. In the Weiner and Walker study the nickel reward was as potent a motivator as was the shock; in Experiment II the anticipated reward had only a small influence on retention. Analysis of the subjects participating in the two experiments provides one possible clue to explain the conflicting results. In the Weiner and Walker study the subjects were paid volunteers, part of a permanent pool of paid subjects at the University of Michigan. Their primary source of motivation to participate in experiments is monetary. In Experiment II the subjects were students enrolled in introductory psychology at the University of Minnesota. Their primary source of motivation to participate in experiments is class credit. It is likely that in this situation money is a more salient and effective motivator for the former than the latter subjects. That is, there may have been an interaction between the type of incentive and the motivations of the subjects.

Experiment III

In the previous two studies the magnitude of the positive incentive was varied. In Experiments III and IV only one monetary reward was employed, while the strength of the aversive shock was manipulated. The four incentive conditions were: win five cents for correctly recalling the stimulus; receive a small shock following incorrect recall; receive a larger shock following incorrect recall; and a control condition. The peak voltage of the smaller shock was in the order of 175 volts, while the larger pulse shock approximated 250 volts. Both shock intensities decayed to near zero after 3 milliseconds.

Inasmuch as differential KOR did not account for the divergent decay rates in the previous experiments, only the feedback conveyed by the shock was used. This procedure minimizes experimenter-subject interactions. Subjects were 24 male students enrolled in introductory psychology at the University of Minnesota.

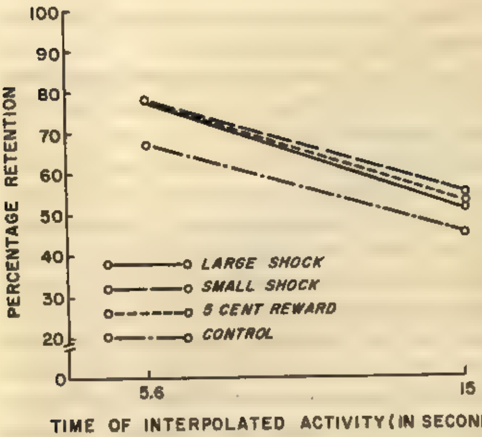

FIG. 3. Percentage retention at two time intervals as a function of incentive, employing two levels of shock intensity.

Results

In Figure 3 the percentage of correct responses over the two time intervals is plotted. An analysis of variance indicates that the main effect does not reach statistical significance, $F(3, 69) = 2.72$, $p < .10$, while the Time × Incentive interaction does not approach significance, $F(3, 69) < 1$. Further analysis reveals that all three motivational conditions differ significantly from the control group ($p < .01$), but not from one another.

Discussion

The results contrast with the previous findings in that the differences between the motivational and nonmotivational conditions at the short time interval are as great as the differences exhibited at the long interval. No definitive conclusions concerning retention can be drawn from this study. However, it is of interest to note that the recall of stimuli in the low shock intensity condition is almost identical with the recall in the high intensity condition.

EXPERIMENT IV

Experiment III was attempted again with some procedural modifications. It was thought that the absence of differences in recall between the shock conditions in Experiment III may have been caused by a failure to discriminate the smaller from the larger shock. In the present study the two intensities were differentially raised: the intensity of the smaller shock was in the order of 200 volts, while the larger shock approximated 300 volts. Also, the retention intervals were altered, and a third recall point was added. The three time intervals were: 1.87 seconds, 7.50 seconds, and 17.0 seconds. Prior to the test trials six practice trials were administered. There were 72 test trials, 6 for each of the 12 (four incentives × three time intervals) experimental conditions. In the first and last 36 trials each of the 12 conditions appeared three times; within each of the 36 trials the order of presentation was randomized. Subjects were 24 male students enrolled in introductory psychology.

Results

In Figure 4 the percentage of correct responses is plotted for the four incentive conditions and the three time intervals. An analysis of variance indicates that there is a significant main effect attributable to the incentive, $F(3, 69) = 4.70$, $p < .01$. The Incentive × Time interaction approaches significance, $F(6, 138) = 1.87$, $p < .10$. A Newman-Keuls paired-mean test was performed to compare the results with earlier findings. The test reveals that there are no significant differences in recall at the short or intermediate time intervals. At the long

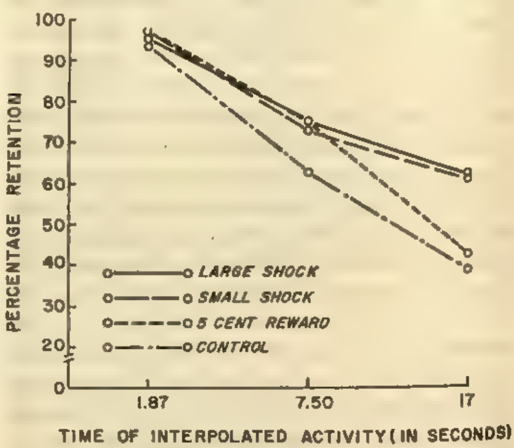

FIG. 4. Percentage retention at three time intervals as a function of incentive, employing two levels of shock intensity.

interval recall of stimuli associated with either the larger or smaller shock is significantly greater than the recall of the control stimuli ($p < .01$) and stimuli linked with a five-cent reward ($p < .05$).

Discussion

As in the prior experiments, retention which was instrumental to the avoidance of shock was enhanced. However, the magnitude of the shock was not related to recall, thus replicating the finding of Experiment III.

The decay rate exhibited by the stimuli paired with a five-cent reward is interesting. At the intermediate time interval these stimuli were retained as well as the stimuli paired with shock, and recall was well under 100%. As the time interval lengthened, the anticipated five-cent reward did not enhance recall, replicating the results of Experiment II. This writer can offer no explanations for some of the undiscussed differences found between Experiment III and Experiment IV (e.g., the differences in recall at the short interval found in Experiment III but not in Experiment IV). It also should be noted that in all conditions the recall at the short time interval approximated 100%. Because prior studies in this series indicated that differential recall was not to be attributed to differences in original learning, no attempt was made to keep recall at the short interval at or below 80%.

The results do indicate another factor which must be specified when investigating the effects of motivation on retention. Previously it was stated that the magnitude and type of incentive in part determine retention. This must now be modified: only in certain cases will the magnitude of the incentive be a relevant variable. That is, there is an interaction between the effects of the magnitude and type of incentive influencing recall.

Experiment V

In Experiments I–IV the motivational manipulation involved the magnitude of positive and negative incentives. The focus of investigation considerably shifts in Experiment V. To account for the differential decay rates exhibited between motivational and nonmotivational conditions, Weiner and Walker (1966) suggested that the greater the motivation during the perception of an event, the less the likelihood that the trace of that event would be subject to retroactive interference. An earlier study by Prentice (1943) had found that differences in retention of material learned under high and low motivational conditions are maximized when subjects are asked to recall following an interfering activity. Consequently, it is hypothesized that motivational factors will have the greatest opportunity of manifesting their influence when interference is maximal. That is, motivational factors are expected to be most efficacious under conditions which maximize forgetting.

In Experiment V, conditions were established to increase the amount of forgetting exhibited in the earlier studies. The incentives in this experiment are the same as those in Experiments I and II, which employed 2 degrees of positive incentive. During the interpolated time interval the subjects were required to read pairs of digits, add them, and state whether the total was odd or even. A metronome striking two beats per second paced this activity. Posner and Rossman (1965) previously demonstrated that this procedure greatly reduces retention.

Other facets of the experimental design were identical with those of Experiment IV. In this and all future experiments there are three recall intervals. Subjects were 16 male students enrolled in introductory psychology.

Results

Figure 5 shows the percentage of retention at the three time intervals for the four incentive conditions. An analysis of variance reveals that there is a significant main effect attributable to the incentive, $F(3, 45) = 15.12$, $p < .01$, and a significant Incentive $\times$ Time interaction, $F(6, 50) = 5.65$, $p < .01$. A Newman-Keuls paired-mean test shows that at the long and intermediate time interval stimuli associated with shock are recalled significantly more often than the control stimuli ($p < .01$). At the intermediate interval stimuli paired

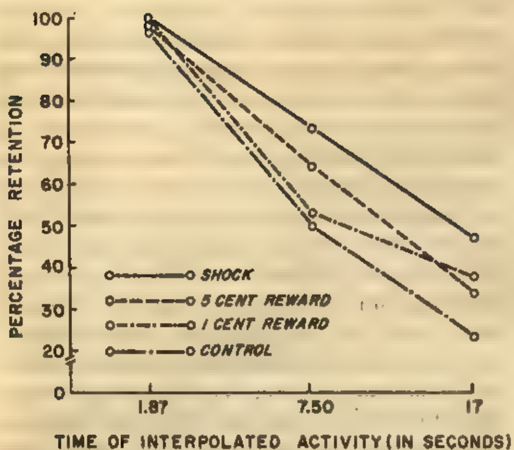

FIG. 5. Percentage retention at three time intervals as a function of incentive, employing 2 degrees of monetary reward and a more difficult interpolated activity.

with shock also are recalled significantly more than stimuli paired with one-cent reward ($p < .01$).

Discussion

For the three identical incentive conditions in Experiment IV and Experiment V the respective total mean recall at the intermediate interval was 71% and 63%; at the long interval recalls were, respectively, 50% and 33%. The interpolated activity in Experiment V clearly had more detrimental effects on retention than the interpolated task employed in Experiment IV. In Experiment V the difference in the percentage of recall between the shock and control stimuli at the long time interval was 22%; in Experiment IV this difference was 16%. The respective differences in retention for the shock versus nickel condition at the long time interval were 15% and 12.5%. At the intermediate time interval the differences between the recall of shock and control stimuli were 24% in Experiment V and 10% in Experiment IV; differences in the recall between shock and nickel stimuli at the intermediate interval were 9% in Experiment V and −1% in Experiment IV. Hence, the difference between the retention of stimuli associated with potent motivational factors as opposed to control stimuli or stimuli associated with a less powerful motivator is greater under conditions which minimize total retention. These results tend to support the hypothesis that motivational factors will be most salient given conditions which maximize forgetting.

This hypothesis, however, is not supported when restricting the data analysis to Experiment V. In that experiment there was no interaction exhibited at the intermediate and long time interval between the shock and control stimuli. That is, the difference in retention between the shock and control stimuli did not increase as the total amount of retention progressively decreased.

A better procedure for the above comparisons would be to conduct two (or more) studies which vary only the difficulty of the interpolated activity. However, the general pattern of results in Experiments I–V do suggest that the influence of motivation on retention is in part a function of the magnitude of incentive, type of incentive, and type (difficulty) of experience intervening between stimulus onset and recall.

EXPERIMENTS VI–VIII

In the preceding five experiments the incentive cue was presented simultaneously with the onset of the stimulus. The temporal sequence of events was:

incentive cue
stimulus
↓
interpolated activity
↓
recall

Therefore, the motivational factors were introduced during the period of trace formation. In the following experiments the temporal locus of the motivational manipulation is altered so that it cannot influence the strength of the original association.

The two additional experimental paradigms in Experiments VI–VIII are:

stimulus
↓
interpolated activity
(*a*) ↓
incentive cue
recall

stimulus
↓
interpolated activity
(b) ↓
incentive cue
↓
recall

Thus, the motivational manipulation is executed during the periods of trace storage and trace retrieval. Introducing the incentive cue simultaneous with stimulus onset or during the interpolated activity theoretically might influence the course of trace decay during the storage period.

Experiment VI

The procedure combines various conditions used in prior experiments. Two degrees of monetary incentive (Experiments I, II, V), three time intervals (Experiments IV, V), and the normal interpolated activity (Experiments I–IV) were used. All stimuli are projected on a blank background. During the recall interval the color cue is projected. The temporal sequence of events therefore is:

stimulus
↓
interpolated activity
↓
incentive cue
recall

Subjects were 16 male students enrolled in introductory psychology.

Results

The results indicate that there are no significant differences between the recall of stimuli as a function of the incentive condition, $F < 1$.

Experiment VII

To increase the amount of motivation to retrieve a stimulus the procedure was modified to include only two conditions. In one condition recall was rewarded with five cents while nonrecall was punished with a shock. In the second condition neither shock nor money was associated with the outcome. Other aspects of the procedure were identical with Experiment VI.

Results

There are again no significant differences between the recall of stimuli in the two motivation conditions, $F < 1$. To indicate the equality of recall, there are 21.4% incorrect responses in the control condition and 23.1% incorrect responses in the incentive condition.

Experiment VIII

It was noted that some subjects tend to emit their responses during the interslide interval, that is, before the onset of the incentive cue. Therefore the cue was brought forward in the temporal sequence of events and presented prior to the recall period:

stimulus
↓
interpolated activity
↓
incentive cue
↓
recall

The cue was on for 7.5 seconds, and the recall period was an additional 6 seconds. There were 60 randomized trials, 10 for each experimental condition (three time intervals × two incentive conditions). All other procedures were identical with Experiment VII.

Results

As in Experiments VI and VII, there are no significant differences in recall as a function of the incentive conditions, $F < 1$.

Discussion

At this point the data indicate that presenting the motivational source during or immediately prior to recall does not enhance retention. Therefore, the effects of motivation on retention are dependent upon the magnitude and type of incentive, nature of the experience intervening between stimulus onset and recall, and the memory process accompanying the motivational input.

A direct comparison between the effects of money and/or shock presented during the time of learning and during the period of retrieval is possible by comparing recall

in Experiment IV with recall in Experiment VI. These studies employed identical shock, nickel, and control conditions; three time intervals for recall; and the same interpolated task. However, the motivational cueing occurred at different places in the memory sequence. (Although Experiment VII combined the shock and nickel conditions, the data were very similar to the results in Experiment VI; Experiment VII therefore is included in the following comparison. Experiment VIII is not considered because the procedural change, unexpectedly, dampened total recall.) In Experiments IV, VI, and VII the recall of stimuli associated with shock was extremely consistent, varying between 76% and 77%. Similarly, recall of stimuli associated with a five-cent reward varied only between 73% and 77%. On the other hand, in Experiment IV there was 67% recall of control stimuli, while in Experiments VI and VII recall of the control stimuli was respectively 78% and 79% (see Figure 6).

Two very different interpretations of the data shown in Figure 6 are offered. It may be that the differences in retention exhibited in Experiments I–IV (cue at onset of stimulus) are not due to an enhanced retention of stimuli associated with positive or negative incentives. Rather the differences in recall are to be attributed to *a relative decrement* in the retention of the

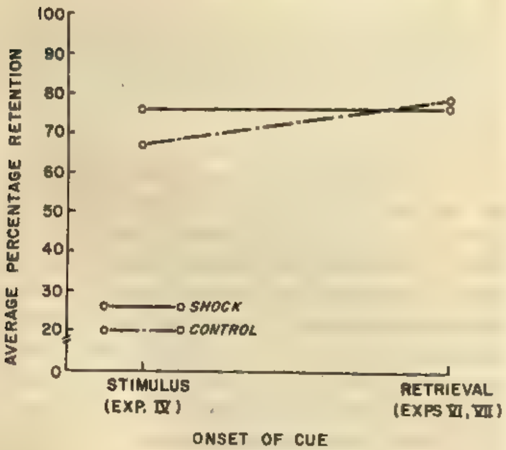

FIG. 6. Average percentage retention of control and shock stimuli in Experiment IV (cue at stimulus onset) and Experiments VI and VII (cue during trace retrieval).

control stimuli. An alternative explanation is that in Experiments VI–VIII the individuals acted *as if all* the stimuli were associated with an incentive until the motivational cue actually was presented. That is, the subjects responded to the onset of the stimulus with heightened motivation because the recall of the stimulus might have a motivational consequence. Given this interpretation, Experiments VI–VIII are somewhat analogous to a situation in which the subject is confronted with a random partial reinforcement schedule. While engaging in the instrumental behavior (storage) it is not possible for the organism to predict the actual outcome; the individual therefore behaves identically in the motivational and "nonmotivational" conditions. This interpretation suggests that there were no "control" stimuli in Experiments VI–VIII. The sequence of events would be portrayed as:

(*a*) Motivational condition
motivational manipulation
↓
stimulus
↓
interpolated task
↓
incentive
recall

(*b*) Control condition
motivational manipulation
↓
stimulus
↓
interpolated task
↓
recall

The conclusion from experiments cueing during recall would continue to be that motivational manipulations during the retrieval process do not enhance recall. However, the possible effect could have been dampened by the sequentially prior expectation of an incentive. This interpretation implies that the differential recall exhibited in Experiments I–V is attributable to an absolute enhancement in the recall of stimuli paired with a motivational variable.

The analysis presented in the above

paragraph indicates that the methodology employed in the prior experiments might be inadequate to ascertain the effects of motivation on memory when cueing occurs during trace retrieval. To investigate this problem a one-trial experiment must be conducted in which the individual is not aware of a potential motivational outcome prior to the period of trace utilization. Experiment IX creates these conditions.

Experiment IX

Subjects were 164 male students enrolled in the introductory psychology course. The experiment was administered on four occasions in one evening to groups ranging in size from 20 to 80. Subjects were asked not to divulge the nature of the experiment to the incoming groups.

A three-page booklet was randomly distributed to the subjects; the first page of the booklet was blank. Subjects silently read the directions printed on page 2 of the booklet. Two groups, produced by two types of instructions, were created: an Intentional Learning group and an Incidental Learning group. Their respective instructions were:

(*a*) You will be seeing a series of words on the screen. Following the list presentation you will be asked to recall the words. Pay close attention to the words as they are flashed.

(*b*) You will be seeing a series of words on the screen. These are practice trials to familiarize you with the experimental procedure which we will be using in the actual experiment which follows.

Prior evidence that motivational factors influencing recall are most effective when forgetting is maximized suggested that there might be an interaction between degree of original learning and subsequent trace retrieval in motivational and nonmotivational conditions. For this reason two groups expected to differ in their original learning were created.

Thirty-six nouns were then projected on a screen. Each stimulus word was visible for 8 seconds, with a .75-second interslide interval. Twelve words were classified as AA on the Thorndike-Lorge (1944) word count, 12 appeared 30–40 times per million, and 12 appeared once per million. The 36 stimuli were presented in a randomized order.

Following the presentation of the stimuli subjects silently read the instructions on the next page of the booklet. Three groups were created by varying the quality and magnitude of the motivation aroused following stimulus presentation. For one group, money was offered as an incentive for stimulus recall; for the second group, achievement motivation was aroused (McClelland, Atkinson, Clark, & Lowell, 1953); a third group was a control group. The respective directions for these groups were:

Now write down the words which were presented on the screen. The order of recall is not important. Just write them down as they occur to you. You can guess if you are uncertain. You will have five minutes for this task.

(*a*) For every word correctly recalled you will win five cents. You can, therefore, win almost $2.00. You will be paid immediately following the experiment.

(*b*) We have found in the past that the ability to remember words is related to general success on exams. So try your best so that your performance reflects your ability.

(*c*) When the five minutes are up, we will collect the booklets. Remain seated during the entire five minutes.

There were six experimental groups (two learning conditions × three retrieval conditions). It was hypothesized that the Intentional Learning group would recall more than the Incidental group, the Achievement and/or Monetary Reward group would recall more stimuli than the control group, and that there would be an interaction between the degree of original learning and the motivational condition at the time of retrieval.

Results

An analysis of variance yielded the expected difference in recall between the groups which differed in their instructions prior to learning, $F(1, 158) = 8.51$, $p < .01$. However, there were no significant differences in recall between the groups which differed in strength of motivation during retrieval, $F < 1$, nor any evidence for the expected interaction, $F < 1$.

Discussion

The results support the previous conclusion that motivational input during the period of trace retrieval does not enhance re-

call. This finding is somewhat disconcerting because other investigators (e.g., Blum, 1961; Bourne, 1955) have established that motivational manipulations at the time of retrieval do affect recall. At the present time the investigator cannot account for this contradiction in results. The great differences between the experiments of Bourne, Blum, and these studies do not permit critical methodological comparisons.

EXPERIMENTS X–XV

There remains the perplexing analysis which suggested that the differences in recall found between conditions in Experiments I–V are not due to an enhancement of the recall of stimuli in the motivational condition. Rather, they seem to be caused by a relative decrement in the recall of the control stimuli. Before considering the theoretical implications of this finding, it is necessary to determine conclusively the facts. Experiments X–XV reveal that this is a more difficult problem than one would anticipate.

EXPERIMENT X

Experiment X was conducted simultaneously with Experiments VI–VIII as part of a master's thesis by Kernoff (1965; reported in Kernoff, Weiner, & Morrison, 1966). The study is not directly related to the problems discussed above, but the findings lead to some procedural changes employed in Experiments XI–XV.

Studies I–V are conceivably subject to some methodological criticisms. Differences in learning in those investigations may have been masked because performance at the short interval was comparatively near asymptote and a relatively insensitive measure of learning was used (cf. Underwood, 1964). Experiment X attempts to replicate the basic findings of Weiner and Walker, employing two methodological changes. First, four-letter consonant stimuli (quadrigrams) rather than trigrams were the to-be-remembered units. The stimuli were formed by adding a consonant to the trigrams. The consonant was not repeated in the trigram, and "v" and "w" were excluded. The stimuli were projected on the screen for 1 second. A second change was instituted to increase the sensitivity of the response indicator. Each response was evaluated on an eight-point scale of approximation to the correct response. Responses were scored one point for each consonant recalled, and two points for each consonant recalled in its correct position. The maximum score of eight was received for perfect recall. There were three recall time intervals: 2.8 seconds, 9.35 seconds, and 17 seconds. Subjects were 20 male students enrolled in introductory psychology.

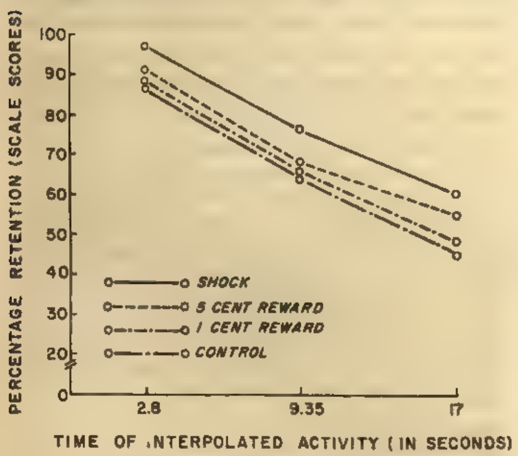

FIG. 7. Percentage retention at three time intervals as a function of incentive, employing 2 degrees of monetary reward, four-letter consonant stimuli (quadrigrams), and a more sensitive response indicator.

Results

The mean response score for each incentive condition at the three time intervals is shown in Figure 7. An analysis of variance reveals a significant main effect due to the incentives, $F(3, 57) = 8.70$, $p < .01$. The Time $\times$ Incentive interaction does not approach significance, $F < 1$. Paired-mean tests were employed to compare these results with the earlier findings. A Newman-Keuls test reveals that there are no significant differences in recall at the short time interval. At the intermediate interval stimuli associated with shock are retained significantly more than the control stimuli ($p < .05$). At the long interval stimuli associated with shock are recalled significantly more than stimuli associated with a

one-cent reward ($p < .05$) and the control stimuli ($p < .01$). Stimuli paired with a five-cent reward also are recalled significantly more than the control stimuli ($p < .05$) at the long time interval.

Discussion

The general pattern of results replicates the findings of Weiner and Walker, save that the effectiveness of the nickel is not as great. The results of this preliminary investigation are essential for the interpretation of the findings in Experiments XI–XV.

EXPERIMENT XI

This experiment directly attacks the issue concerning the absolute versus relative effects of motivation on retention. In the previous studies in this series a "differential" experimental method has been used (Lawson, 1957). That is, each subject served as his own control and received multiple stimulus presentations. This is a within-subjects experimental design. A second possible procedure is known as the "absolute" method (Lawson, 1957). In that procedure different subjects are used in different experimental conditions; this is a between-subjects design.

The two different experimental approaches have yielded disparate results in psychological research. A number of studies within the domain of motivation report an expected behavioral difference as a function of motivation when the differential method is employed, but the findings have not been replicated when using an absolute method. Lawson (1957) found that "visual discrimination performance varies with incentive amount only when *S*s have experience with different amounts in association with different stimuli [p. 39]." Pubols (1960), summarizing the research on incentives, concluded that incentives affect learning when training is by the differential rather than by the absolute method. In two recent studies of human learning, Harley (1965a, 1965b) has confirmed Pubols' conclusions. Similarly, Grice and Hunter (1964), in an article appropriately entitled "Stimulus intensity effects depend upon the type of experimental design," found that "substantially greater (signal intensity) effects are obtained if individual *S*s are exposed to the different intensities than if each *S* experiences only one intensity value [p. 247]." Further, Wright (1965) was able to establish a learned hunger drive when a within-subjects experimental design was used, whereas numerous previous investigators were not able to find this result using between subject comparisons.

One method to determine the relative versus absolute effect of motivation on memory is to compare results of studies using the two procedures outlined above. Grice and Hunter (1964) state that "this turns out to be a reasonable, but neglected, experimental design in psychological research [p. 248]." If the recall of control stimuli in the absolute method exceeds the recall of control stimuli in the differential method, then the differential procedure results in a decrement in the recall of those stimuli. Motivation would then relatively but not absolutely enhance recall. On the other hand, if the recall of the control stimuli in the absolute condition equals the recall in the differential condition, then the recall of stimuli associated with incentives must have been absolutely enhanced in Experiments I–V. This analysis can be expressed somewhat differently to include the recall of stimuli associated with a motivational factor in an absolute procedure. If the differential method was effective because there was a decrement in the recall of the control stimuli, then in the absolute procedure there should be no difference between the recall of control and motivational stimuli. Conversely, if motivation absolutely enhances retention, then in the between-subjects design the recall of stimuli associated with an incentive should exceed the recall of the absolute control stimuli.

In Experiment XI the conditions necessary to test these comparisons are established. Some adjustments are made in the experimental procedure. The previous studies consistently have shown no differences in recall between the penny and control condition. The smaller monetary reward is therefore not adding any information. In the remaining studies there are only three

experimental conditions: shock, five-cents reward, and a control condition. The stimuli in Study XI were the quadrigrams used in Experiment X. Also, feedback was conveyed following correct recall in the five-cent condition. A pleasant bell was sounded to signal a monetary reward.

Subjects were 57 male students enrolled in the introductory psychology course at the University of California, Los Angeles. Eighteen subjects were tested using the customary differential method; 18 were in an absolute control group, and 21 were in the absolute shock incentive group. Establishing the latter condition created a number of difficulties. If there is a potential shock on every trial, then the total number of shocks received in the absolute and differential procedures would differ. This could result in some adaptation and a loss of the motivational effectiveness of the shock in the absolute method. It is possible to equate the potential number of shocks by giving only 24 trials in the absolute procedure, inasmuch as ⅓ of the 72 trials are cued for shock in the differential method. However, this would confound the amount of practice and proactive inhibition between the two methods. An alternative procedure is to convey to the subject that on a randomly selected ⅓ of the trials he may receive shock if incorrect. This equates the number of potential shocks in the between-subjects and within-subjects methods. This procedure was used in the between-subjects design. The trials associated with shock were isomorphic with the potential shock trials in the differential method.

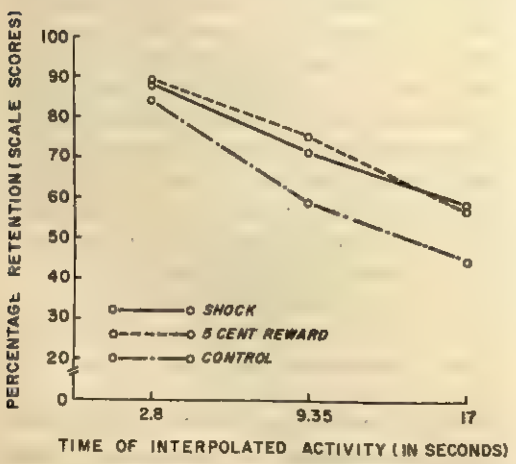

FIG. 8. Percentage retention at three time intervals as a function of incentive, employing quadrigrams as stimuli and the differential procedure.

There is no absolute five-cents reward condition. The findings for the shock condition are sufficient to test the hypotheses under consideration.

Results

In Figure 8 the percentage retention scale scores for the three conditions in the differential procedure are plotted. The data replicate the finding that stimuli associated with a motivational outcome are retained significantly more than the control stimuli. In contrast to our previous studies, the nickel again is as potent a determinant of retention as shock. An analysis of variance on this data indicates a significant main effect due to the incentive, $F(2, 34) = 13.29$, $p < .01$. The Incentive × Time interaction approaches significance, $F(4, 68) = 2.06$, $p < .10$. A Newman-Keuls test reveals that there are no significant differences in recall at the short time interval. At the medium and longer interval recall of the nickel and shock stimuli is significantly greater ($p < .01$) than the recall of the control stimuli.

In the absolute or between-subjects comparison respective recall at the short, medium and long intervals in the shock condition is 89%, 75%, and 61%. In the control condition the respective recall is 87%, 72%, and 61%. There clearly is a minimal difference in recall between the two conditions.

Table 1 compares the total recall in the within-subjects and between-subjects procedures. The table reveals that the recalls of the control stimuli in the absolute method (hereafter referred to as the straight control stimuli) is greater than the recall of the control stimuli in the differential procedure, and approximates the recall of the shock stimuli in the differential procedure.

Discussion

The results seem to confirm the previous suspicion that the retention of stimuli associated with motivational factors is not absolutely enhanced. Rather, there appears to be a decrement in the recall of stimuli

TABLE 1

Summary of Results, Experiments XI–XV, with Studies Grouped According to Type of Stimuli

Experiment	Stimuli	Subjects	Onset of Cue	Procedure: Differential N	Differential Control[a]	Differential Shock[a]	Absolute N	Absolute Control[a]	Absolute N	Absolute Shock[a]
XI	Quadrigrams	Male	Stimulus	18	63	72	18	74	21	75
XV	Quadrigrams	Male	Interpolated Activity	18	65	70	18	71		
XII	Trigrams	Male	Stimulus	18	80	89	18	78	21	93
XIII	Trigrams	Male	Interpolated Activity	18	81	88	18	77		
XIV	Trigrams	Female	Interpolated Activity	18	74	85	18	77		

[a] Percentage retention in scale scores.

coupled with a nonmotivational condition in the differential procedure. The results also substantiate the general conclusions of Pubols and Grice and Hunter that between-subjects designs are less likely to yield an expected motivational effect than a within-subjects design. Further discussion of the results of this investigation and the remaining studies will be postponed until the entire series of investigations (XI–XV) is presented.

Experiment XII

The findings in Experiment XI were of sufficient import to warrant a replication. One change was made in the experimental procedure: trigrams rather than quadrigrams are the to-be-remembered stimuli. In retrospect, the exact reason for this modification is somewhat obscure. The best guess is that it was decided to reinstate some of the conditions used in the original Weiner and Walker study. Subjects were 57 male students enrolled in introductory psychology. Eighteen subjects were tested with the differential method. There were 18 subjects in the absolute control condition, and 21 in the absolute shock condition.

Results

Figure 9 portrays the results when the differential method was employed. As expected, there is a significant main effect due to the incentive, $F(2, 34) = 17.38$, $p < .01$. The Time × Incentive interaction also is significant, $F(4, 68) = 5.10$, $p < .01$. A Newman-Keuls test shows that there are no significant differences in retention at the short interval. At the intermediate interval stimuli cued for shock are retained significantly more than the control stimuli, ($p < .01$). At the longer interval both shock stimuli and stimuli cued for five cents are recalled significantly more than the control stimuli, ($p < .01$).

In the between-subjects design recall at the three time intervals in the shock condition is 99%, 92%, and 87%. For the absolute control group recall at the three intervals is 97%, 74%, and 65%. As in the differential method, recall of motivational stimuli is

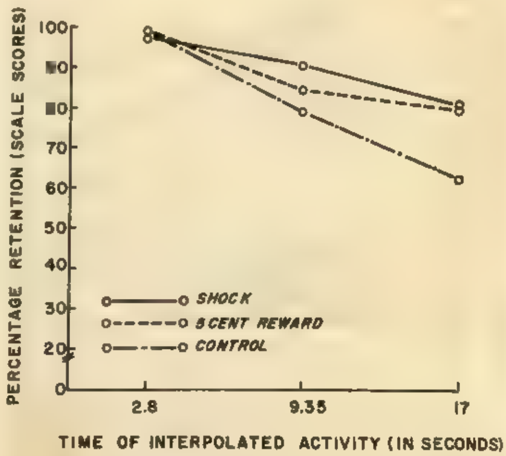

Fig. 9. Percentage retention at three time intervals as a function of incentive, employing trigrams as stimuli and the differential procedure.

greater than the recall of control stimuli ($t = 4.32$, $df = 37$, $p < .01$).

Table 1 shows the total recall of shock and control stimuli in the within-subjects and between-subjects procedures for Experiment XII. It is clear from the table that the two procedures yield virtually identical results. In both conditions there is a significant difference in recall between the motivational and nonmotivational stimuli.

Discussion

The results certainly were surprising. The data indicate that there is an absolute enhancement of the retention of stimuli associated with a motivational factor. The results and inferences of Experiment XI therefore are not substantiated.

The only purposive change between Experiments XI and XII was that Experiment XI employed four-letter consonants as stimuli, while in Experiment XII the stimuli were three-letter consonants.

EXPERIMENT XIII

It was essential to attempt to replicate the apparently conflicting results of the prior two studies. In Experiment XIII trigrams were again the to-be-remembered units (as in Experiment XII). Two changes were made in the experimental design. First, there was no absolute shock condition. The significant comparisons in the prior two studies concerned the recall of control stimuli in the absolute and differential procedures. Hence only an absolute control condition is included in this experiment. Secondly, the motivational cue was presented at the *onset* of the interpolated activity, rather than at the onset of the stimulus. One long-term goal of this research program is to determine the place in the memory sequence at which the motivational input will maximally affect retention. Earlier studies in this series have indicated that motivational input at the time of stimulus presentation does influence recall, but this is not true when the input occurs during the period of retrieval. In this study the motivational cue appears after the offset of the stimulus, during the period of trace storage. Further implications of this change will be discussed later in the paper. While it is indeed risky to vary a factor when attempting to replicate a finding, the author decided to take this risk because of the multitude of problems which needed exploration. If the experiment replicated the prior results, in spite of the induced change, then two findings would emerge from one experiment.

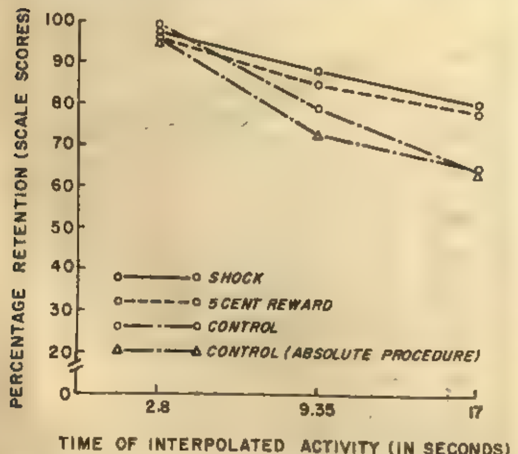

FIG. 10. Percentage retention at three time intervals as a function of incentive, employing trigrams as stimuli, the absolute and differential procedures, and cueing for motivation at the onset of the interpolated activity.

Eighteen male students were tested with the differential procedure, and 18 were in the absolute control condition.

Results

Figure 10 illustrates the recall in the three conditions tested with the differential procedure and the recall of the straight control stimuli. Again with the differential procedure there is a significant main effect attributable to the incentive, $F(2, 34) = 11.77$, $p < .01$, and a significant Time × Incentive interaction, $F(4, 68) = 5.73$, $p < .01$. Recall of the control stimuli does not differ from the recall of the motivational stimuli at the short time interval. At the medium interval the shock and nickel stimuli are recalled significantly more than the control stimuli, ($p < .05$). The differences in recall also are exhibited at the long interval, ($p < .01$). The recall of the absolute control stimuli is virtually identical to the recall of the control stimuli in the differential procedure. The respective total

recall in the differential and absolute procedures is given in Table 1.

Discussion

The results replicate the findings of Experiment XII. As Table 1 shows, the pattern of results in Experiments XII and XIII is almost identical. The data consistently demonstrate that the retention of trigrams associated with a potential shock is absolutely enhanced.

EXPERIMENT XIV

To be absolutely certain about the reliability of the findings in Experiments XII and XIII, the experiment was conducted again. In Experiment XIV there is one procedural change: subjects are females. In all the prior studies the subjects were males. The use of males in the first study by Weiner and Walker was entirely chance; males happened to be available in the subject pool. After the initial finding the experimenter was somewhat reluctant to include female subjects because sex differences pervade so many problem areas in psychology. An unreported pilot study conducted earlier in this series did reveal that females were behaving differently than males in situations employing "right" and "wrong" feedback, but were replicating the male results when only shock feedback was used. It was therefore decided not to use females in the ensuing experiments. The major impetus for Experiment XIV was the alarming availability of female subjects in introductory psychology and a scarcity of male subjects. The experimental procedure was identical with that in Experiment XIII.

Results

The pattern of recall at all time intervals is virtually identical with the recall in the previous study. In the differential method there is a main effect attributable to the incentive conditions, $F(2, 34) = 11.50$, $p < .01$. The Time × Incentive interaction does not reach significance, $F(4, 68) = 1.70$, $p < .25$. There is the familiar difference in recall at the medium and long interval between the shock and nickel versus control stimuli ($p < .01$). The amount of recall in the straight control condition is very similar to recall of the control stimuli in the differential procedure. Table 1 gives the total percentage recall in the absolute and differential procedures.

Discussion

Experiments XII and XIII were replicated. The total recall for females was slightly lower than that of males, but the general results are strikingly consistent (see Table 1).

EXPERIMENT XV

Experiments XII–XIV are convincing; with trigrams as stimuli there is an absolute facilitation in retention in the motivational conditions. Experiment XV reverts to the use of quadrigrams as stimuli. The procedure is identical with that used in Experiment XIII, which used male subjects.

Results

Figure 11 gives the retention for subjects in both experimental methods. In the differential procedure there is the significant main effect attributed to the incentive, $F(2, 34) = 3.54$, $p < .05$. The Time × Incentive interaction approaches significance, $F(4, 68) = 2.33$, $p < .10$. There is no sig-

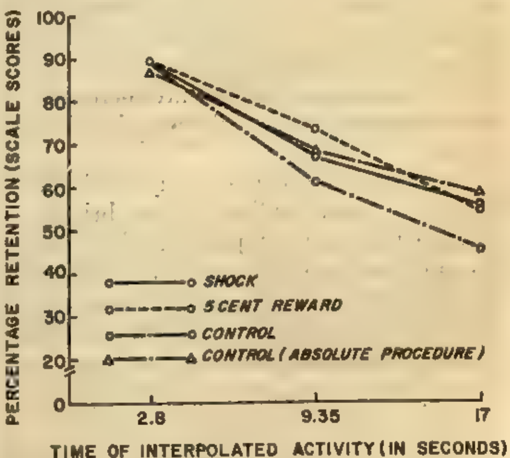

FIG. 11. Percentage retention at three time intervals as a function of incentive, employing quadrigrams as stimuli, the absolute and differential procedures, and cueing for motivation at the onset of the interpolated activity.

nificant difference in recall between the motivational and control stimuli at the short time interval. At the medium interval there is a significant difference between the recall of the nickel and control stimuli, $p < .01$. At the long interval both the shock and nickel stimuli are recalled significantly more often than the control stimuli, $p < .05$. The amount of recall in the absolute condition again exceeds the recall of the control stimuli in the differential procedure (see Table 1).

Discussion

The pattern of results replicates Experiment XI. With quadrigrams as stimuli there seems to be a relative decrement in the recall of control stimuli, rather than an enhancement in the retention of stimuli associated with motivational factors.

Reconsideration of Experiments XI and XII

The results in Experiments XI and XII have been replicated in three studies, and at this time must be considered reliable. Table 1 presents the total recall of shock and control stimuli in Experiments XI–XV. Using quadrigrams as stimuli, the table indicates that the differences in recall in the differential method appear to be caused by a relative decrement in the retention of the control stimuli. Employing trigrams as stimuli, Table 1 reveals that the difference in recall in the differential method seems to be attributable to an absolute increment in the retention of stimuli associated with shock.

Before considering some general implications of these findings, let us briefly return to the data comparison which led to the questioning of the relative facilitation in the retention of the to-be-shocked stimuli (see Figure 6). To recapitulate briefly, Figure 6 seems to indicate that there is a decrement in the recall of control stimuli cued for incentives during the period of stimulus onset. This was inferred from the data in Experiments VI–VIII, which found no difference in the retention of stimuli cued during the period of retrieval. It was then asked whether there was a decrement in the recall of the control stimuli, or whether cueing at retrieval rendered all the stimuli motivationally relevant. Experiments XI–XV provide evidence which favors the latter alternative. In the studies employing trigrams as stimuli (Experiments XII–XIV), there is an absolute increment in the recall of the potential shock stimuli. Experiments I–VIII employed trigrams as stimuli. Hence in those experiments there also must have been a facilitation in the recall of the shock stimuli. The equality in the recall of both shock and control stimuli when cueing during retrieval, and the equivalence of that recall with the recall of the to-be-shocked stimuli cued during stimulus onset, indicate that when cueing during retrieval the subject reacts to all the stimuli with heightened motivation.

There are other, undoubtedly more significant, implications of the results in Experiments XI–XV. The differential procedure yielded the same findings when either trigrams or quadrigrams were the stimuli. However, the addition of the absolute procedure demonstrates a difference in the retention function between the two types of stimuli. The between-subjects design led to the differentiation of the results obtained with the within-subjects design. Stated somewhat differently, Experiment XI (quadrigrams as stimuli) yielded different results with different experimental procedures, while Experiment XII (trigrams as stimuli) yielded the same results across the different procedures. Thus, there is an interaction between the type of experimental design and the type of stimulus material. This interaction indicates that different experimental interpretations are needed when trigrams and quadrigrams are the to-be-remembered materials. More generally, the interaction suggests that research in psychology might profit from an experimental design similar to the multitrait, multimethod approach advocated by Campbell and Fiske (1959) in correlational research. Multimethod comparisons perhaps are a necessary part of psychological research.

Why is it that when trigrams are the to-be-remembered stimuli motivation absolutely enhances retention, while this does

not appear to be the case when quadrigrams are the stimuli? At this time the author can only vaguely speculate about some of the dimensions which distinguish the stimuli. First, there is absolutely greater retention of trigrams than quadrigrams. It was previously suggested that the effects of motivation on retention are in part of a function of the total amount of forgetting. The results of Experiments XI and XII do not support the hypothesis that motivational factors are most effective when forgetting is maximal. However, the general relationship between motivation and total forgetting might be a determinant of the strange pattern of results exhibited in Experiments XI–XV. A second difference between trigrams and quadrigrams is their degree of meaningfulness. Trigrams are likely to be more meaningful than quadrigrams, although specific associative norms on quadrigrams remain to be collected. Other research in the area of motivation and memory has indicated that meaningfulness is an important determinant of retention. White, Fox, and Harris (1940) found that during a hypnotic trance the recall of meaningful material was facilitated. This was not true for the recall of nonsense syllables. Rosenthal (1944), in a more complex study, also demonstrated that the recall of meaningful material is enhanced if retrieval is during a hypnotic state. Meaningfulness may be an important dimension in the present research, and conceivably could be a significant factor differentiating trigrams from quadrigrams. Still a third difference between the stimuli may be that trigrams are more likely to be retained with the aid of special associative or mnemonic devices. For example, one subject reported that the stimulus XGP was easy to remember because he formerly was an aspiring general practioner. Such strategies perhaps are less possible with quadrigrams; this is an unmeasured dimension in the present research.

It is also of interest to note the resurgence of the potency of the five-cent reward. It is regretful that the nature of the research program makes some conclusions very tentative. It may be that the nickel was effective because the pleasant signal accompanied the correct response. However, this feedback was used in a new population, and the nature of the subject population may have been the significant variable.

In summary, the variables which effect the motivation and memory linkage tentatively include: the magnitude of incentive, the type of incentive, the nature of the experiences intervening between stimulus onset and recall, the point in the memory sequence at which the motivation is introduced, the nature of the feedback, the stimulus material, the type of experimental design, and complex interactions between these variables.

GENERAL ISSUES

Rehearsal

In the earlier paper of Weiner and Walker the following statement was made concerning rehearsal:

> It might be argued that the interaction between the time interval and the incentive conditions was mediated by differential rehearsal of the stimuli. It is conceivable that the greater the incentive value of the stimulus, the greater is the tendency of the subjects to repeat that stimulus. The differential rehearsal hypothesis is especially provocative because it is generally accepted that learning increases as a function of the number of repetitions of the to-be-learned material. If motivational manipulations result in differential rehearsal, then the degree of learning becomes confounded with the storage process, and demonstrating that motivation influences retention certainly would be a formidable problem. In this experiment subjects were paced during the interpolated task to minimize the amount of rehearsal; there is no evidence that subjects do or do not covertly rehearse one set of stimuli more than another set. [p. 192]

There are now a number of strong arguments against the differential rehearsal explanation of the results. First, the speed and difficulty of the interpolated task limits the feasibility of this explanation. In addition, rehearsal would be most likely to occur during the .70-second interslide interval between the offset of the stimulus and the onset of the interpolated activity. Therefore, cueing at the onset of the interpolated activity (Experiments XIII–XV) should lessen the differences in recall between the motivational and nonmotiva-

tional conditions as compared with studies which cue at the onset of the stimulus. The data, however, do not support this supposition. Further, it was found that a more difficult interpolated activity increased the differences in recall between the motivational and control stimuli. This would not be expected if differential rehearsal caused the differences in retention. Rehearsal should be less likely to occur as the difficulty of the interpolated task increases. Finally, the differential rehearsal explanation would not shed any light upon the complex interaction found between the absolute and differential procedures and the stimulus material.

Repression

Previous research (e.g., Clemes, 1964; Russell, 1952; Zeller, 1952) has suggested that events associated with unpleasant affective states tend to be "repressed"; that is, the traces of those events are unavailable for immediate retrieval. Many of the experimental studies of repression are methodologically inadequate (cf. Weiner, 1966). However, there are some conclusive experimental demonstrations of the phenomenon (e.g., Clemes, 1964), and a great wealth of clinical observations (Freud, 1946) supporting the concept of repression. Conversely, the studies reported in this monograph demonstrate that the retention of material associated with unpleasant events is enhanced, rather than dampened.

A critical difference between the investigations presented here and previous research and observations concerning repressed material is that in the present procedure retention is instrumental to the avoidance of a potential shock. Repression is conceptualized as an ego function (Freud, 1936) and is regulated in accordance with the pleasure-pain principle. Within the framework of analytic theory it therefore is quite conceivable that the retention of events associated with unpleasant outcomes will be facilitated, given proper circumstances. In other discussions of repression, "forgetting" rather than retention is the more adaptive psychological process; in the present paradigm the reverse is true. This analysis is similar to Dulany's (1957) view that both perceptual vigilance and perceptual defense can be exhibited, depending on the nature of the response instrumentalities.

Action Decrement

Walker (1958) has presented a general theory of learning and retention which interrelates concepts of arousal, action decrement, and consolidation. The major prediction of his theory is that high arousal during learning makes the trace of the material less available for immediate recall, but results in greater permanent memory. These predictions have been substantiated (Kleinsmith & Kaplan, 1963; Walker & Tarte 1963).

In the present studies, Walker's predictions are not confirmed. Stimuli a priori considered to be highly arousing because of their association with an affective consequence are more likely to be recalled after a relatively short time interval than stimuli considered to be relatively low in arousal value. At this time the author cannot reconcile the results supporting Walker's theory with the data from the present series of studies. The experiments are quite different, and methodological comparisons are not possible. However, the instrumentalities of retention in the present studies again may be a crucial difference. Walker considers the decrement in the availability of highly arousing stimuli to be an adaptive process which ultimately strengthens the stimulus trace. In the present investigations clearly the more adaptive procedure is to have the material immediately available for recall.

SUMMARY

Fifteen studies were presented which investigate the effects of motivation on memory. In a variation of the Peterson and Peterson technique devised to study short-term retention, stimuli (trigrams) were cued for various incentives. For some stimuli recall was rewarded with money, while for other stimuli nonrecall was punished with a shock. In Experiments I–IV the magnitude of the aversive shock and monetary reward was varied. Recall was enhanced in the motivational conditions when

compared to a control condition, although the magnitude of the potential shock was not related to recall. In these studies there were no differences in recall at a short time interval, and recall at that point was approximately 80%. Further, there were differential decay rates for identical stimuli. Therefore, differences in recall were attributed to differences in retention (storage) rather than to differences in the degree of original learning. Experiment V demonstrated that the differences in recall between conditions is maximized as the task interpolated between stimulus onset and recall becomes more difficult.

In Studies I–V the onset of the stimulus and the motivational cue were presented simultaneously. In Experiments VI–IX the cue was presented immediately prior to or during the period of trace retrieval. There were no differences in recall between the conditions in these experiments.

Experiments X–XV examine whether the differences in retention exhibited in Experiments I–V are to be attributed to a decrement in the retention of the control stimuli, or to an increment in the retention of the motivational stimuli. To investigate this problem it is necessary to employ both between-*S*s and within-*S*s experimental designs. The stimuli were trigrams or quadrigrams (four-letter consonants). The findings appear to indicate that with trigrams as stimuli there is an absolute enhancement of the recall of stimuli associated with a motivational state. However, with quadrigrams as stimuli the differences in retention seem to be caused by a decrement in the recall of the control stimuli. These findings were consistent in five experimental investigations. The results were replicated when the stimuli were cued either during stimulus onset or during the onset of the interpolated activity. Possible reasons for the unexpected pattern of results were discussed.

The article concludes with a discussion of the relevance of this work to repression and action decrement. It also was concluded that differences in recall between motivational and nonmotivational conditions were not caused by differential rehearsal of the stimuli.

REFERENCES

ATKINSON, J. W. *An introduction to motivation.* Princeton: Van Nostrand, 1964.

BLUM, G. S. *A model of the mind.* New York: Wiley, 1961.

BOURNE, L. E., JR. An evaluation of the effects of induced tension on performance. *Journal of Experimental Psychology,* 1955, **49,** 418–422.

CAMERON, D. C. Remembering. *Nervous and Mental Disease Monograph,* Whole No. 22, 1947.

CAMPBELL, D. T., & FISKE, D. W. Convergent and discriminant validation by the multitrait, multimethod matrix. *Psychological Bulletin,* 1959, **56,** 81–105.

CARON, A. J. & WALLACH, M. A. Recall of interrupted tasks under stress: A phenomenon of memory or learning? *Journal of Abnormal and Social Psychology,* 1957, **55,** 372–381.

CLEMES, S. R. Repression and hypnotic amnesia. *Journal of Abnormal and Social Psychology,* 1964, **69,** 62–69.

DULANEY, D. E., JR. Avoidance learning of perceptual defense and vigilance. *Journal of Abnormal and Social Psychology,* 1957, **55,** 333–338.

FITZGERALD, D., & AUSUBEL, D. P. Cognitive versus affective factors in the learning and retention of controversial material. *Journal of Educational Psychology,* 1963, **54,** 73–84.

FREUD, A. *The ego and the mechanisms of defense.* New York: International Universities Press, 1946.

FREUD, S. *The problem of anxiety.* New York: Norton, 1936.

GRICE, G. R., & HUNTER, J. J. Stimulus intensity effects depend upon the type of experimental design. *Psychological Review,* 1964, **71,** 247–256.

HARLEY, W. F., JR. The effect of monetary incentive in paired associate learning using a differential method. *Psychonomic Science,* 1965, **2,** 377–378 (a).

HARLEY, W. F., JR. The effect of monetary incentive in paired associate learning using an absolute method. *Psychonomic Science,* 1965, **3,** 141–142 (b).

HEYER, A. W., JR., & O'KELLY, L. I. Studies in motivation and retention: II. Retention of nonsense syllables learned under different degrees of motivation. *Journal of Psychology,* 1949, **27,** 143–152.

HULL, C. L. *A behavior system.* New Haven: Yale University Press, 1952.

KEPPEL, G. Problems of method in the study of short-term memory. *Psychological Bulletin,* 1965, **63,** 1–13.

KERNOFF, P. Affect and short-term retention. Unpublished master's thesis, University of Minnesota, 1965.

KERNOFF, P., WEINER, B., & MORRISON, M. Affect and short-term retention. *Psychonomic Science,* 1966, **4,** 75–6.

KLEINSMITH, L. J., & KAPLAN, S. Paired associates learning as a function of arousal and interpolated interval. *Journal of Experimental Psychology,* 1963, **65,** 190–193.

LAWSON, R. Brightness discrimination performance and secondary reward strength as a function of primary reward amount. *Journal of Comparative and Physiological Psychology,* 1957, **50,** 35–39.

LEVINE, J. M., & MURPHY, G. The learning and forgetting of controversial material. *Journal of Abnormal and Social Psychology,* 1943, **38,** 507–512.

LEWIN, K. *The conceptual representation and the measurement of psychological forces.* Durham, N. C.: Duke University Press, 1938.

MCCLELLAND, D. C., ATKINSON, J. W., CLARK, R. A., & LOWELL, E. L. *The achievement motive.* New York: Appleton-Century-Crofts, 1953.

MELTON, A. W. Implications of short-term memory for a general theory of memory. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 1–12.

MELTZER, H. The present status of experimental studies on the relationship of feeling to memory. *Psychological Review,* 1930, **37,** 124–139.

PETERSON, L. R., & PETERSON, M. Short-term retention of individual verbal items. *Journal of Experimental Psychology,* 1959, **58,** 193–198.

PETERSON, L. R., PETERSON, M. J., & MILLER, A. G. Short-term retention and meaningfulness. *Canadian Journal of Psychology,* 1961, **15,** 143–147.

POSNER, M. I., & ROSSMAN, E. The effect of size and location of information transforms upon short-term retention. *Journal of Experimental Psychology,* 1965, **70,** 496–505.

PRENTICE, W. C. H. Retroactive inhibition and the motivation of learning. *American Journal of Psychology,* 1943, **56,** 282–291.

PUBOLS, B. H., JR. Incentive magnitude, learning, and performance in animals. *Psychological Bulletin,* 1960, **57,** 89–115.

RAPAPORT, D. *Emotions and memory.* Baltimore: Williams & Wilkins, 1942.

ROSENTHAL, B. G. Hypnotic recall of material learned under anxiety- and non-anxiety-producing conditions. *Journal of Experimental Psychology,* 1944, **34,** 369–389.

RUSSELL, W. A. Retention of verbal material as a function of motivating instructions and experimentally-induced failure. *Journal of Experimental Psychology,* 1952, **43,** 207–216.

SPENCE, K. W. *Behavior theory and conditioning.* New Haven: Yale University Press, 1956.

THORNDIKE, E. L., & LORGE, I. *The teachers word book of 30,000 words.* New York: Teachers College, Columbia University, 1944.

UNDERWOOD, B. J. Speed of learning and amount retained: A consideration of methodology. *Psychological Bulletin,* 1954, **51,** 226–282.

UNDERWOOD, B. J. Degree of learning and the measurement of forgetting. *Journal of Verbal Learning and Verbal Behavior,* 1964, **3,** 112–129.

WALKER, E. L. Action decrement and its relation to learning. *Psychological Review,* 1958, **65,** 129–142.

WALKER, E. L., & TARTE, R. D. Memory storage as a function of arousal and time with homogeneous and heterogeneous lists. *Journal of Verbal Learning and Verbal Behavior,* 1963, **2,** 113–119.

WEINER, B. The effects of motivation on the availability and retrieval of memory traces. *Psychological Bulletin,* 1966, **65,** 24–37.

WEINER, B., & WALKER, E. L. Motivational factors in short-term retention. *Journal of Experimental Psychology,* 1966, **71,** 190–193.

WHITE, R., FOX, G., & HARRIS, W. Hypnotic hypermnesia for recently learned material. *Journal of Abnormal and Social Psychology,* 1940, **35,** 88–104.

WITMER, L. R. The associative-value of three-place consonant syllables. *Journal of Genetic Psychology,* 1935, **47,** 337–360.

WRIGHT, J. H. Test for a learned drive based on the hunger drive. *Journal of Experimental Psychology,* 1965, **70,** 580–584.

ZELLER, A. F. An experimental analogue of repression: III. The effect of induced failure and success on memory measured by recall. *Journal of Experimental Psychology,* 1952, **42,** 32–38.

(Received April 26, 1966)

Vol. 80, No. 19 Whole No. 627, 1966

Psychological Monographs: General and Applied

INTROSPECTIONIST AND BEHAVIORIST INTERPRETATIONS OF RATIO SCALES OF PERCEPTUAL MAGNITUDES[1]

C. WADE SAVAGE[2]

University of California, Los Angeles

(a) The psychological magnitudes involved in perception are observable but private according to the introspectionist, public and nonobservable according to the behaviorist. (b) In their constructions of psychophysical scales, both the introspectionist and the behaviorist rely on a principle of correspondence between psychological magnitude and O's estimates of physical magnitude. The former bases this principle on hypotheses concerning O's perceptual mechanism; the latter regards the principle as a stipulative definition. (c) Psychophysical laws obtained by ratio and other scaling procedures are explanations of O's behavior on the introspectionist view, descriptions of O's behavior on the behaviorist view. In the past the introspectionist and the behaviorist interpretations have been run together, thus making it possible to amalgamate the advantages and obscure the deficiencies in both. These deficiencies suggest that the concept of psychological magnitude ought to be abandoned, and that perceptual psychophysical scaling procedures should be regarded as procedures for measuring perceptual abilities. This suggestion violates both the old psychophysics of Fechner and the new psychophysics of Stevens. But it dissolves the traditional question of whether psychophysical measurement is possible, as well as the contemporary question of whether the validity of competing psychophysical scales can be determined. The above analysis is illustrated with a hypothetical fractionation experiment in which a ratio scale of psychological length is constructed.

PSYCHOPHYSICS is generally regarded as the attempt to measure or scale psychological magnitudes. Very roughly we may distinguish jnd (confusion, discriminability) scales, partition (category, equisection) scales, and ratio (magnitude) scales of psychological magnitudes. The "new" psychophysics, for which S. S. Stevens is largely responsible, holds that the ratio scale is most desirable, for the following reasons. It is superior to a jnd scale, since it is constructed by a "direct" method (Stevens, 1958a, p. 387), and since jnd's are not "subjectively equal" on so-called prothetic continua (Stevens & Davis, 1936, pp. 411–416, and Stevens, 1954, pp. 30–31; 1957, pp. 154, 172; 1960b, p. 227; 1960c, pp. 57–59; Hirsh, 1952, pp. 10–11). It is superior to a partition scale, since it enables us to say not only that one psychological entity is greater than another, but also how much greater (Stevens & Davis, 1936, pp. 406–407, and Stevens, 1959b, p. 611; 1960a, p. 28; 1960b, pp. 228–230); and also because the ratio scale contains the partition scale (1960c, pp. 53–54).

Furthermore, ratio-scaling procedures have led to the discovery of a psychophysical law of great generality and theoretical power (Stevens, 1957, p. 162; 1958b, pp. 192–194; 1960a, pp. 28–29; 1960b, pp. 234–235; 1961, p. 84; 1962, pp. 30–32). Stated as a first approximation the law is

$$\Psi = k\Phi^n, \qquad (1)$$

where Φ is the stimulus magnitude in physical units, Ψ is the psychological magnitude

[1] After the acceptance of this *Monograph*, I sent the manuscript to S. S. Stevens who has prepared the appended reaction (pp. 33–38). Considerations of time precluded my sending Stevens' response to Savage for further comment. *Editor*.

[2] This research was brought to completion with the bibliographical assistance of James R. Shaw, whose services were provided by a University of California research grant.

in psychological units, k is a constant determined by the choice of units, and n varies according to the sense modality in question. (1) is a power law and contrasts with Fechner's logarithmic law which states that

$$\Psi = kn \log \Phi, \quad (2)$$

where Ψ is the psychological magnitude measured in jnd units. The "new" psychophysics claims that (2) is invalid, because the method for obtaining it is "indirect" and employs jnd's as psychological units. (1), on the other hand, is said to be based on "direct" methods and consequently to employ psychological units which accurately represent psychological magnitudes. (For a review of recent developments in psychophysical scaling see Ekman & Sjöberg, 1965.)

Modern psychophysicists usually maintain or imply that they have cast off the dualist metaphysics and the introspectionist methodology which frustrated their predecessors. Thus Galanter (1962, pp. 92–93) says:

> The name psychophysics derives from the classical question about the relation between the physical environment and the mind. Today, modern psychophysicists are not professionally concerned with this philosophical issue of the mind-body relation, but rather with the constraints that are placed upon the behavior of a person in his judgments, actions, and so on, by the sea of physical energies that surround him.

And in Hirsh (1952, pp. 15–16) we read:

> The influence of behaviorism in American psychology is easily seen in modern psychophysics. We no longer look for relations between stimuli and sensations but rather relations between stimuli and responses. We can observe responses, the elements of behavior, and measure them, whereas the private sensation, which remains as untouchable as it was in Fechner's day, does not concern us. We do not ask whether or not a man hears a tone. We seek only to find whether or not he responds in a specified way to a tone. We can have measurement, then, on both sides of the psychophysical relation. The "psycho" part refers merely to behavior.

(We will not discuss the interesting historical question of whether Fechner was trying to measure private, introspectional sensation, as is commonly supposed. The reader is referred to Boring [1928] and Johnson [1929] who take opposite sides on this question.)

The major claim of this study is that such declarations of philosophical enlightenment are premature. Many of the "new" psychophysicists officially subscribe to a behaviorist and operationist philosophy, but in their experimental and theoretical work employ introspectionistic assumptions and lapse into an introspectionistic view of the nature of psychophysical measurement. They assure us that theirs is an attempt to measure behavioral responses, and yet they employ methods whose rationale seems to be that they enable the experimenter to quantify those private sensations which were formerly regarded as directly inaccessible. This indictment could be completely substantiated only by considering a large number of scaling methods and the work of a large number of psychophysicists. In this study we will concentrate on ratio-scaling methods and on the work of S. S. Stevens, who is the principal architect of the "new" psychophysics.

The principal feature of the analysis in this paper is a distinction between introspectionist and behaviorist interpretations of perceptual magnitudes, the psychological magnitudes involved in perception. The two interpretations are sketched in an early section of the paper and then presented in detail and criticized in succeeding sections. In the final section we illustrate and discuss the unfortunate consequences of the almost universal failure to distinguish the two interpretations. Our analysis may indicate that the concept of psychological magnitude is illegitimate, and that the attempt to scale such magnitudes ought to be abandoned. This would be to abandon not only the orientation of Stevens' "new" psychophysics, but also that of the "old," which Fechner founded. Even if this more sweeping conclusion cannot be sustained, our analysis at least shows that neither ratio, partition, nor jnd scales can be properly assessed and compared without distinguishing the introspectionist from the behaviorist view of psychophysical measurement.

A Sample Ratio-Scaling Experiment

Ratio scales can be constructed either by numerical estimation methods—magnitude production and magnitude estimation, or by fractionation methods—ratio production and

ratio estimation. (For an exhaustive classification and description of scaling methods see Stevens, 1958b.) The analysis of this study will be based on a sample experiment in which a scale for psychological length is constructed by the method of ratio estimation. Our major results apply equally, however, to ratio scales constructed by any of the other appropriate methods. The experiment presented here is imaginary and stylized. Nonetheless, its main features have been extracted from actual experiments, if not for length then for similar continua. And our analysis can, with only trivial adjustments, be applied to any actual ratio scaling experiment. We will call our sample experiment "M," for "mak scale."

It consists of two parts. In each trial of the first part the experimenter (E) presents the person experimented on (O) with a standard rod, Φ_s, and asks him to select a comparison rod, Φ_c, which looks one-fourth as long. In each trial of the second part O is asked to select comparison rods which look one-half as long as the standards. The data thus obtained are represented by the solid lines in Figure 1. In both fractionations O consistently overestimates the comparison rod (or—should we say?—underestimates the standard). The equation for the lower line is

$$\Phi_c = .33\,\Phi_s, \tag{3}$$

for the upper line,

$$\Phi_c = .58\,\Phi_s. \tag{4}$$

Notice that the numerals along each axis of Figure 1 represent physical rather than psychological magnitudes. To measure the psychological magnitude involved in O's perception, a unit of psychological length must be chosen, and the data on which Figure 1 is based restated in terms of that unit.

The choice of a unit of measurement is made primarily on the basis of convenience. Since it is convenient, let us stipulate that the psychological length associated with a stimulus rod of 100 cm. is 100 psychological units, and let us call the unit thus defined the *mak*. (For the origin of both the unit and its name see Reese, Reese, Volkmann & Corbin, 1953, p. 41.) This choice of unit determines

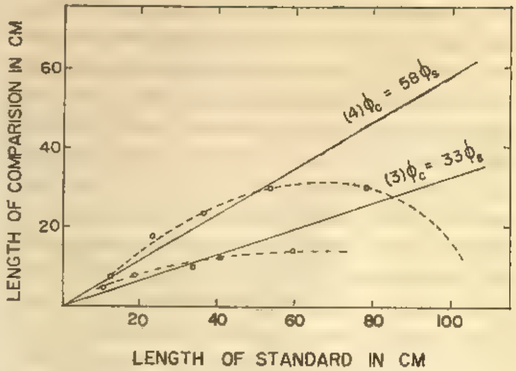

FIG. 1. A plot of hypothetical data for two length fractionation experiments. The upper solid line represents O's visual estimates of one-half, the lower solid line O's visual estimates of one-fourth. The dash lines are derived by procedures and for purposes which are explained in a later section.

the first point in a plot of psychological length against physical length, the point marked "A" in Figure 2. Referring back to Figure 1, we find that O would estimate a 58-cm. rod to be half as long as a 100-cm. rod. Hence, the psychological magnitude associated with the former must be half that associated with the latter. Since the former was 100 maks, the latter must be 50 maks.

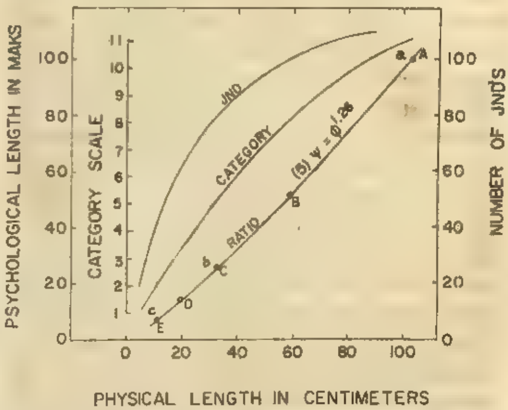

FIG. 2. Three psychological curves for length estimates. The lower, ratio curve is obtained from the solid lines of Fig. 1. Points A–E are obtained from the upper solid line of Fig. 1, points a–c from the lower solid line in Fig. 1. The middle curve represents the data of an imaginary category scaling experiment, the upper curve the data of an imaginary jnd experiment. The relations of these two curves to the ratio curve are discussed in later sections.

Thus we obtain the point labeled "B." Referring again to Figure 1, we see that O would estimate a 33.6-cm. rod to be half as long as a 58-cm. rod. Hence, the psychological magnitude associated with the former must be 25 maks, half that associated with the former. This gives us Point C. D, E, and any further points desired are obtained by the same method, and a curve is fitted visually to these points.

The equation for this curve is

$$(5) \qquad \Psi = \Phi^{1.26},$$

where Ψ is the psychological magnitude expressed in maks and Φ is the physical magnitude expressed in centimeters. A Ψ-Φ curve can also be constructed in a similar manner from Line (3), by assuming that when O says one rod looks one-fourth as long as a second the psychological magnitudes associated with the two rods stand in the ratio 1:4. We thus obtain Points a, b, and c, which precisely correspond to Points A, C, and E of the previous construction. The two psychophysical curves are thus found to coincide, which shows, presumably, that O's one-half and one-fourth estimates of length are made on the same psychological continuum.

The other two curves in Figure 2 are concave downward. The lower of these is constructed from data obtained in an imaginary category scaling experiment, in which O is asked to assign numerals from "1" to "11" to rod lengths. The O is asked to assign the numerals in such a way that the difference between a rod to which "1" is assigned and a rod to which "2" is assigned is the same as the difference between a rod to which "2" is assigned and a rod to which "3" is assigned, and so on. The category scale is represented on the inside left-hand ordinate. The upper curve in Figure 2 is constructed from data obtained in an imaginary jnd experiment. Number of jnd's is represented on the right-hand ordinate. Although all three curves are based on imaginary data, they are nevertheless typical of the results of actual experiments dealing with length, area, finger span, and so on (see Reese, Reese, Volkmann, & Corbin, 1953; Stevens & Galanter, 1957; Stevens & Stone, 1959).

TWO INTERPRETATIONS

(α) It is obvious that any interpretation of Experiment M must contain a *distinction between psychological* (subjective, apparent) *magnitudes and physical* (objective, real) *magnitudes* (see Guilford, 1954, p. 21). The physical entities Φ_1, Φ_2, Φ_3, etc.—of which physical length is a magnitude—are associated with psychological entities, Ψ_1, Ψ_2, Ψ_3, etc.—of which psychological length is a magnitude, and which are involved in O's perception of physical entities (identity of subscript indicates association of physical entity with psychological entity).

Magnitudes may be classified as intensive or extensive (Bergmann & Spence, 1944, pp. 5–8; Savage, 1963, Ch. 3; Stevens & Volkmann, 1940, pp. 330–331). If psychological length is an intensive magnitude, then any Ψ can meaningfully be said to be greater than, equal to, or less than another. Consequently, we can ask whether psychological length varies with physical length and, if so, whether inversely or directly. If psychological length is an extensive magnitude, then any given Ψ can meaningfully be said to be half as great, twice as great, 10 times as great, etc. as another. Consequently, we may sensibly attempt to measure psychological length in the fullest sense, that is, to erect a ratio scale for this magnitude, and to determine precisely how, in terms of its scale and the centimeter scale, psychological length varies with physical length. Experiment M attempts to do this, and psychophysical law (5) is the result.

(β) Any interpretation of Experiment M must commit the procedure therein employed to a *principle of correspondence* between relative psychological length and O's estimates of relative physical length. In the experiment (*a*) E measures the physical length of rods to be presented to O, (*b*) O estimates the relative physical length of these rods, (*c*) E represents the data thus obtained in Figure 1, and (*d*) E scales psychological length by constructing Figure 2 from Figure 1. (*a*), (*b*), and (*c*) are straightforward enough, but (*d*) is puzzling. How, from the data of *physical* measurement and *physical* estimates, is a psychological magnitude

scaled? Such scaling will appear to be a conjuring trick unless certain of its assumptions are brought to light. Chief among these is what we have called the principle of correspondence, the general statement of which is: If O estimates that Φ_1, Φ_2, Φ_3, etc. stand in the physical relation R, then Ψ_1, Ψ_2, Ψ_3, etc. stand in the psychological relation R.

The instance of this principle which is used in constructing a jnd scale is (A): If O estimates that Φ_1 and Φ_2 are just noticeably different and that Φ_2 and Φ_3 are jnd, then the interval between Ψ_1 and Ψ_2 is equal to that between Ψ_2 and Ψ_3. The instance used in constructing a partition scale is (B): If O estimates that the interval between Φ_1 and Φ_2 equals that between Φ_2 and Φ_3, then the interval between Ψ_1 and Ψ_2 equals that between Ψ_2 and Ψ_3. The instance used in constructing a ratio scale is (C): If O estimates that Φ_1 and Φ_2 stand in the same ratio as Φ_2 and Φ_3, then Ψ_1 and Ψ_2 stand in the same ratio as Ψ_2 and Ψ_3. The more specific instance used in constructing the ratio scale in Experiment M is (Cl): If O estimates that Φ_1 and Φ_2 stand in the ratio $m:n$, then Ψ_1 and Ψ_2 stand in the ratio $m:n$. Notice that the method of numerical estimation also employs a principle of correspondence, one that is intimately related to (C). The principle is (D): If O assigns to Φ_1, Φ_2, Φ_3, etc. the numerals m, n, o, etc. respectively, then $\Psi_1 = k(m/n)\ \Psi_2$, $\Psi_2 = k(n/o)\ \Psi_3$, etc., where either $k = 1$ or $k \neq 1$. If we make the standard assumption that $k = 1$, then the principle becomes (D1): If O assigns to Φ_1, Φ_2, Φ_3, etc. the numericals m, n, o, etc. respectively, then Ψ_1, Ψ_2, Ψ_3, etc. stand in the ratios $m:n:o:$ etc. (For examples of the explicit use of (Cl) see Harper & Stevens, 1948, p. 345, and Reese, 1943, pp. 22–23).

(γ1) The introspectionist interpretation of Experiment M holds that *Ψs and their magnitudes are privately observable and that the principle of correspondence embodies a theory of* O*'s perceptual mechanism*. This interpretation preserves a simple, attractive, and familiar view of the nature of psychophysics. Psychological entities are held to occupy a private realm distinct from the public realm occupied by physical entities. Nonetheless, psychological magnitudes are as "empirically real" as and no less fundamental than physical magnitudes. Consequently, it is a legitimate scientific enterprise to try to measure Ψ magnitudes, much as we measure Φ magnitudes (e.g., with a meter stick), and to attempt to discover the mathematical relation between Ψ magnitudes and Φ magnitudes, just as we attempt to discover the mathematical relation between two Φ magnitudes. Psychophysics is thus "an exact theory of the functionally dependent relations of body and soul" (Fechner, 1966), the attempt to extend the best procedures of physical science into the psychological realm. Another advantage is that the introspectionist interpretation leaves no doubt that M is a *psychological* experiment. How do the *physical* measurements and *physical* estimates obtained in M become transformed into a scale of *psychological* length? The introspectionist answers that O perceives psychological entities by some inner sense and uses these inner estimates to make outer estimates. Although O reports only the latter, E can make use of the former by hypothesizing a certain perceptual mechanism in O.

This hypothesis, however, produces one of the disadvantages in the introspectionist interpretation. The introspectionist assumes that O visually estimates the length of rods, Φs, by introspecting the magnitude of psychological entities, Ψs. But it is doubtful that any perceptual mechanism of this sort is at work in O when he estimates rod length. Putting the difficulty differently, principle of correspondence (Cl) is based on premises describing O's perceptual mechanism which are entirely problematic. These premises will be stated and thoroughly examined in the sequel. A related disadvantage is that the introspectionist interpretation does not seem to fit the phenomenological facts of Experiment M. O observes rods and estimates their relative length. But there is no phenomonological evidence that he observes private, psychological entities and estimates their magnitude. A further disadvantage is that Ψs are held to be private, hidden from the view of everyone but O. E must therefore rely on

O's unconfirmed Ψ estimates in constructing a ratio scale of psychological magnitude. This seems to place psychological magnitudes outside the pale of "objective" scientific investigation and measurement.

(γ2) The behaviorist interpretation holds that Ψs *and their magnitudes are nonobservable* and that *the principle of correspondence is true by definition*. This interpretation is consistent with the phenomenological fact that Experiment M seems to involve observations of physical entities only. A further advantage is that psychological entities are no longer located in some private realm, directly accessible only to O. Now they can be regarded as generally available to scientists, like any other magnitude which is capable of metricization and scientific treatment. But these advantages are purchased at what appears to be a price.

For the distinction between psychological and physical magnitudes loses its sharpness and obviousness. It is now seen as unwise to think of Experiment M as an attempt to "discover the relation between mind and body," since the dualism of parallel realms implicit in such a characterization is being brought to question. The dualist view implies that psychological entities are, although private, as "empirically real" as physical entities. But the behaviorist holds that they are theoretical constructs or "fictions," conceptual inventions of the experimenter devised for some anticipated scientific use. The principle of correspondence, (Cl), is the tool for building these constructs. It is a stipulative definition, not a description of some mechanism underlying O's perception of physical length. And the psychophysical law, (5), made possible by these conceptual maneuvers must be regarded, not as an *explanation*, but as a *description* of O's length perceptions. Ψs thus become a useful but dispensable *façon de parler*.

The Introspectionist Interpretation

The Nature of Psychological Entities

One difficulty for the introspectionist interpretation arises from its assertion that O observes private entities of which psychological length is a magnitude. The O *sees* rods, walls, and the experimenter; he *feels* his chair and other physical objects; he *hears* the sound of E's voice; and so on. This description of what O observes does not mention any private, psychological entities. If O does observe such entities, how does he do it: by seeing them, smelling them, "intuiting" them? This and related difficulties cannot be assessed until more content is given the notion of a psychological entity. Several suggestions deserve consideration.

The *first* is that Ψs are visual sensations: private mental events or processes which occur within O during rod perception. Thus briefly stated, the suggestion remains obscure, since we have not yet given any instances of visual sensations or any reason to believe they exist. The obscurity can be removed by offering visual afterimages as familiar paradigm examples of visual sensations. On this suggestion, to say of O that he has a visual sensation is to say that he has a visual afterimage or something like one. A clear meaning is thus given to the assertion that visual sensations are private. The O's afterimages are *logically* private, since it is a logical, not merely an empirical fact, that his afterimages can be perceived only by him. Furthermore, the mode of perception now becomes clear: visual afterimages and whatever is like them are *seen*.

A *second* suggestion distinguishes between psychological and physical magnitudes, but not between psychological and physical entities. The psychological length of rods is their length as perceived by O; their physical length is their length as measured by E. Ψ_1 and Φ_1 refer, not to different entities, but to different aspects of the same entity. As it stands, this suggestion is unclear, since we still do not know what "length as perceived by O" means. One way of removing the unclarity is to suggest that perceived length is the length in his visual field of rods seen by O. (Other ways of removing the unclarity take us into the behaviorist interpretation.) This suggestion preserves the logical privacy of psychological *magnitudes*, since it is a logical fact that only O can perceive the length in his own visual field of a rod. It does not preserve the privacy (nor the separateness) of psychological *entities*. In addition, the suggestion implies that psychological length is perceived by sight.

A *third* suggestion holds that Ψ magnitudes are magnitudes of physiological processes occurring within *O* during rod perception. These may be specified as retinal processes (length or area of retinal stimulation, retinal electrical potential, etc.); as optic nerve processes (frequency of nerve impulses, number of activated fibres, etc.); or as brain processes (area of stimulation in the occipital lobes, electrical potential in the lobes, etc.). It is not as strange as it may seem to classify this suggestion as introspectionist. For it contends, like the other two, that Ψ magnitudes are privately observed by *O*. Unlike the others, it does not make clear by what faculty Ψs are perceived. Are optic nerve impulses *felt*? Are retinal processes *seen*? Furthermore, physiological processes, unlike visual sensations, are contingently rather than logically private. Although it may in fact be true that only *O* perceives the retinal processes occurring to him, these *can* be observed by other perceivers, either now or in the future, by means of suitable instruments.

Each of the three suggestions above must be rejected for the same two reasons.

First, there is no phenomenological evidence whatsoever that *O* observes during M any of the Ψ magnitudes suggested. As regards the first suggestion, no afterimages are induced in *O*; he sees none of these, nor anything like them. The complainant who says that an afterimage is a poor paradigm for a visual sensation must produce a better one, on pain of leaving the notion of a visual sensation wholly obscure. In defense of the second suggestion, it may be said that any rod seen by *O* must have a length in his visual field. Even if we concede this less than clear contention, still there is no reason to believe that *O* is aware of every—or even of any—rod's length in his visual field. He makes no reports of and seems to pay no attention to such magnitudes. From the objector who complains that length in the visual field is a poor paradigm for the notion of apparent length, we must demand a better one, else the notion of apparent length will remain entirely obscure. As for the third suggestion, one way of emphasizing that *O* perceives no retinal, nerve, or brain processes is by pointing out that an observer who had never heard of retinas, optic nerves, or brains could function quite as well in Experiment M as a professional physiologist. Our general criticism in this paragraph can be reinforced by contrasting the sort of instructions actually given in M ("Select the rod which is half as long as the one I now hold up") with the instructions which would be required in any of the suggestions mentioned (e.g., "Select the rod which has a size in your visual field half that of the rod I now hold up"). The new instructions would produce entirely different experiments.

Secondly, Experiment M cannot be construed as a procedure for measuring the Ψs mentioned in any of the three suggestions. The standard procedures for measuring the size of an afterimage, or the size in the visual field of a rod, require *O* to view the image or rod against a screen at a fixed distance and and to indicate the area on this screen occluded by the image or rod. The *E* can then express the size of the afterimage, or the size in *O*'s visual field of the rod, in terms of the size in centimeters or inches of the occluded area. No screen is used in Experiment M, and no determinations of occluded area are made. It is even more obvious that M is not a procedure for measuring physiological processes. The area of retinal stimulation is measured by applying an electroretinoscope, or, less directly, by computations based on the construction of the eye and the laws of optics. Instrumental and computational methods are also employed in measuring nerve impulses, and brain processes. None of these methods is found in M.

The foregoing analysis shows that the concept of a psychological entity is unclear and scientifically unacceptable. When we try to understand the concept by producing possible instances—afterimages, things in the visual field, physiological processes—the two objections just presented become applicable. These objections are so conclusive, and in a way so obvious, as to make it seem that psychological entities could not be any of the things suggested. Thus the concept dissolves like a mist exposed to the light of day. And it is a mist, a vague penumbra, in the thinking of the introspectionist. He has some vague notion of psychological entities—

"sensations," he usually calls them—located in a private mental realm within the perceiver. But they are not afterimages, nor things in the visual field, nor physiological processes, all of which are measurable. Regarded in this obscure fashion, "sensations" are probably unintelligible, and certainly outside the reach of scientific investigation and measurement. The *O*'s sensations cannot be measured by *E* because of their privacy. But neither can they be measured by *O*, since they are not the sort of entities to which the operations required in measurement can be applied. Measurement may be defined as *a procedure for assigning numerals to a class of objects by an operation of comparing the objects with a unit or units*. For physical length the operation of comparison is laying a ruler alongside the object, for physical weight it is placing the object on a balance. What operation of comparison with a unit can *O* apply to his sensations?

Faced with this problem, the introspectionist may concede that the mak scale does not qualify as sensation measurement, on the above definition of that term, but insist that the definition is too narrow. Measurement, he may say, is *any assignment of numerals to a class of objects which represents the magnitude ratios of the objects*. This definition tells us that it is the result and not the manner of numeral assignment which is important. Numerals may be assigned by means of an operation of comparing objects with a unit, or in some quite different manner, like that in Experiment M. As long as the result is an assignment which represents magnitude ratios, measurement can be said to have taken place.

The content of this reply was suggested to the writer by Stevens' discussions of the nature of measurement. Stevens (1959a, p. 24) maintains that measurement is "the assignment of numerals to aspects of objects or events according to rule." On the basis of four different rules for numeral assignment he distinguishes four types of scales: nominal, ordinal, interval, and ratio (Stevens, 1951, 1959a). But the rules mentioned describe only the *result* of numeral assignment. Thus the rule for an interval scale is: Assign numerals so as to represent equal intervals. The rule for a ratio scale is: Assign numerals so as to represent equal ratios. Stevens (1951, pp. 28–29) denies that a physical operation of addition is required to create even a ratio scale. And he (Stevens, 1959b, pp. 614–615) says that Fechner's mistake was in believing that "measurement must be reducible to counting [the constituents of sensation]." All this seems to suggest that *any* method which produces a ratio scale (or any of the other types) is properly understood as measurement, which is just to say that it is the result and not the manner of a procedure of numeral assignment which makes it one of measurement. (It should be noted that Stevens officially subscribes to a behaviorist position and that he would probably deny vehemently that his theory of measurement suggests an introspectionist interpretation of psychological scaling. Nevertheless, see the extensive discussion of his writings in a later section.)

Whether this reply can or cannot be forced on Stevens, it is important to formulate it and to see that it is unacceptable, unacceptable because it confuses sensation *estimation* with sensation *measurement*, and illegitimately substitutes the one for the other. An analogy is helpful in presenting this critical distinction. Suppose that *O* sees from a distance the shadows cast by rods on a wall, but that *E* can neither see nor apply his meter stick to the shadows. Wishing nevertheless to measure them, and believing that *O* is able to make accurate estimates of shadow length, *E* uses him in a two-part experiment. In each trial of the first part he asks *O* to locate the rod, Φ_1, which casts a shadow, Ψ_1, one-fourth as long as the shadow, Ψ_2, cast by rod Φ_2. In each trial of the second part he asks *O* to designate the rod which casts a shadow one-half as long as the shadow cast by a given rod. With these data in hand *E* constructs first a figure like Figure 1 and then one like Figure 2, and announces with elation that he has measured the inaccessible shadows.

It is clear, however, that shadows have not been measured. The shadow observer determines shadow length, not by measurement, but by direct estimation. The shadow experimenter determines shadow length by relying on the observer's estimates, that is to say, by *in*direct estimation. Similarly,

the "sensation" observer determines the magnitude of his sensations, not by measurement, but by direct estimation. And the sensation experimenter determines the magnitude of O's sensations by relying on the latter's sensation estimates, that is to say, by *in*direct estimation. To call any of these procedures measurement is to violate a distinction which must be preserved by any acceptable definition of the term, the distinction between estimates and measurements. (For more on this distinction, see Savage, 1963, pp. 160–162, 250–252, 300–302.)

The Principle of Correspondence

On the introspectionist interpretation this principle is derived from certain hypotheses about the nature of O's perceptual mechanism. We will illustrate the derivation for the principle employed in obtaining Line (5) from Line (4).

(Ca) If O (indirectly) estimates that Φ_1 and Φ_2 stand in the ratio 1:2, then O (directly) estimates that Ψ_1 and Ψ_2 stand in the ratio 1:2;

(Cb) If O (directly) estimates that Ψ_1 and Ψ_2 stand in the ratio 1:2, then Ψ_1 and Ψ_2 stand in the ratio 1:2;

(Cl) Hence, if O (indirectly) estimates that Φ_1 and Φ_2 stand in the ratio 1:2, then Ψ_1 and Ψ_2 stand in the ratio 1:2.

Informally, (Ca) says that O's estimates of psychological ratios are the same as his estimates of associated physical ratios, (Cb) asserts that O's estimates of psychological ratios are accurate, and (Cl) concludes that psychological ratios correspond to estimated physical ratios. The reader may find it helpful in understanding this argument and the perceptual mechanism it describes, to employ the shadow analogy presented earlier, that is, to think of Ψs as shadows observed only by O and Φs as rods which cast the shadows.

Premise (Cb). This premise is often assumed, rarely justified. Stevens assumes it in a number of places (1936, p. 408; 1956, p. 18), and so do Guilford and Dingman (1954, p. 395). At times Stevens seems to be attempting a justification of the premise (1951, pp. 40–41; 1954, p. 30; 1956, pp. 2, 23). Let us ask how we may determine whether O's psychological estimates are accurate. It would seem that the method must be analogous to that for determining whether estimates of physical ratios are accurate. Suppose we wish to know whether O's estimate that a given rod is half as long as another is accurate. We simply measure the physical lengths of the two rods and then compare measured physical length with estimated physical length. If the ratio as determined by measurement is the same as the ratio as estimated by O, then O's estimate is accurate; otherwise it is inaccurate. Note that the rods must be measured by some procedure which does not depend on O's estimates of rod lengths. Otherwise the accuracy test will be circular and not a genuine test.

By analogy, a test of accuracy for O's estimates of psychological length seems to require a procedure for measuring psychological length. This procedure must be applied and then measured psychological length compared with estimated psychological length. But the procedure for measuring psychological length must be independent of O's estimates of psychological length, if the test is to be noncircular. The mak-scale procedure is thus precluded, since it relies on O's psychological estimates. But what other procedure is available? If, as it appears, there is none, then (Cb) is unverifiable. Let us recapitulate and broaden the difficulty. Experiment M is put forward by its introspectionist proponents as a method for measuring psychological length. Now the experiment rests on the assumption that O's estimates of psychological length are accurate. But this assumption cannot be verified without some method for measuring psychological length which is prior to and independent of M. Hence, to recommend M as a method for measuring psychological length begs the question of whether psychological length is measurable.

Since physiological processes are only contingently, and not logically, private, the above objection does not apply unmodified to the suggestion that O estimates physiological processes in M. Where O's estimates of psychological length are construed, for instance, as estimates of nerve-impulse frequency, we *can* test their accuracy in a noncircular manner by connecting an oscilloscope to electrodes placed on the nerve in question. Even so, related difficulties arise.

First, O's (putative) nerve-impulse estimates have not been tested for accuracy. So the assumption of their accuracy is, although not unverifiable, unverified. Secondly, in testing O's (putative) nerve-impulse estimates for accuracy—which must be done before the mak scale can be accepted—we measure with instruments the very magnitude which M is supposed to scale. M is therefore superfluous, since in justifying it we accomplish its purpose. Thirdly, M cannot be construed as a method for *measuring*, as opposed to estimating, nerve-impulse frequency. The only possible justification for employing M to quantify a physiological magnitude is that the magnitude is presently inaccessible to existing instruments and physiological techniques, and that we may scale the magnitude by means of O's inner observations of it until improved instrumentation and technique make reliance on such a second-best method unnecessary. But this justification clearly implies that Experiment M provides us only with a method for *estimating* a physiological magnitude (allegedly) observed by O and that measurement of such magnitudes is accomplished by instruments and physiological techniques. The importance of the distinction between estimation and measurement has already been sufficiently emphasized.

Premise (Ca). This premise says that O's estimates of psychological ratios are the same as his estimates of the associated physical ratios. But what is its basis? Remember that O provides E with estimates of *physical* magnitudes only in Experiment M. He does not say "Ψ_1 is half as great as Ψ_2," but rather "The length of this rod (Φ_1) is half as great as the length of that rod (Φ_2)." Why assume that when O makes this latter report Ψ_1 and Ψ_2 stand in the ratio 1:2? Perhaps he has learned to say "Φ_1 is half as great as Φ_2" when Ψ_1 and Ψ_2 stand in the ratio 1:3 (Garner [1954, p. 74] is the only experimenter known to the writer who seems to consider such possibilities). If so, and if his Ψ estimates are accurate, then the mak scale in Figure 2 does not correctly represent Ψ ratios. For, on this assumption, Point B should be placed at 33.3 on the y axis, Point C at 11.1, and Point D at 3.7. This would would produce a different scale of psychological magnitude, and a different psychophysical law.

We must know precisely how O's internal estimates relate to his external estimates in order to obtain the correct curve in Figure 2. But no one possesses this knowledge at present, and there is no clear way of obtaining it. The introspectionist holds that Ψ magnitudes are private to the observer. How then can E discover which Ψ estimates serve as a basis for O's Φ estimates of one-half? It seems that the only conclusive way is by asking O: "When you estimate that Φ_1 and Φ_2 stand in the ratio 1:2, what estimate do you make of the relative magnitude of Ψ_1 and Ψ_2?" To see that the meaning of this question is not clear we need only to phrase it in accordance with any of the suggestions regarding the nature of psychological entities mentioned earlier; that is, "When you estimate that a given rod is half as long as another, what estimate do you make of your afterimages (sizes in your visual field, areas of stimulation on your retina)?" Sophisticated observers, as well as naïve ones, would be at a complete loss in the face of such questions.

It is important to realize that neither intraobserver agreement nor interobserver agreement can be used to determine the truth or falsity or (Cl). There is an inclination to suppose that if several observers estimate that Φ_1 and Φ_2 stand in the ratio $m{:}n$, then Ψ_1 and Ψ_2 stand in that ratio. There is an even stronger inclination to suppose that if a single observer estimates on several occasions that Φ_1 and Φ_2 stand in the ratio $m{:}n$, then Ψ_1 and Ψ_2 stand in that ratio. But how are these assumptions to be justified? It is possible for several observers to estimate that Φ_1 and Φ_2 stand in the ratio $m{:}n$ when Ψ_1 and Ψ_2 stand in the ratio $m{:}n$ for one observer, $n{:}o$ for another, $p{:}r$ for a third, and so on. It is possible for a single observer to estimate that Φ_1 and Φ_2 stand in the ratio $m{:}n$ when Ψ_1 and Ψ_2 stand in the ratio $m{:}n$ on one occasion, $n{:}o$ on another, $p{:}r$ on a third; and so on. These possibilities can be ruled out only by comparing estimated physical ratios with actual psychological ratios. Such comparison is possible only if there is some way of determining psychological ratios which is independent of the

observer's estimates. No method of this type is available. Nor is one possible if psychological entities are logically private. For if they are, then every method for determining their actual magnitude must rely on the observer's estimates of their magnitude, and is therefore viciously circular.

It will be said that misleading estimates like those imagined above occur only where the observer attends to the stimulus rather than the sensation, that is, where he commits the "stimulus error", and that they can be avoided by giving observers instructions to attend only to the sensation. But how can *E* be certain that *O* has followed these instructions? Again, it is impossible without an adequate, noncircular method for comparing psychological magnitudes with estimated physical magnitude. Whatever the instructions to the observer, intraobserver and interobserver agreement provide no better grounds for thinking that *O*'s estimates accurately reflect psychological length ratios than they do for thinking that his estimates accurately reflect physical length ratios. Consequently, statements like the following are erroneous:

> In scaling experiments we are forced to assume the uncontrolled ability of the subject accurately to report his sensations . . . the reproducibility of the data upon repetition of [the] experiment lends some support to this assumption [Galanter, 1962, p. 142].

The Purpose of Ratio Scaling

What, on the introspectionist interpretation, is the purpose of constructing the mak scale of psychological length? To some the answer may seem obvious. Many, if not all, empirical scientific laws are statements relating two or more variables which are defined independently of one another. And those laws which give us the mathematical relation are obviously more useful than those which do not. For instance

$$V = kT \tag{6}$$

is the physical law relating the volume and the temperature of a gas where pressure is held constant. No further justification is needed for measuring pressure and temperature than that it makes the formulation and verification of (6) possible. And the attempt to discover (6) needs no justification. Similarly, Psychophysical Law (5) relates two variables, except that one of these is psychological. (5) tells us how psychological length varies with physical length when the experimenter's instructions to the observer as well as certain other elements of the perceptual situation are held constant. Measuring physical length, by a meter stick, and psychological length, by the mak scale, are adequately justified by pointing out that they make the discovery of (5) possible.

The above argument is dubious, first of all, in its contention that the attempt to discover correlations between two or more variables needs no justification. Surely some correlations are more useful than others, and surely some are completely useless. But let this pass. The more important criticism for our purpose is that psychological length does not appear to be a magnitude like pressure or temperature. Pressure and temperature are observable, "empirically real" magnitudes, whereas psychological length does not appear to be so. In addition, pressure is defined independently of temperature, and vice versa; but psychological length does not appear to be definable independently of physical length. These points will become clear in the discussion of the behaviorist interpretation to follow.

A second introspectionist justification for measuring psychological length points out that (Ca)-(Cb)-(Cl) is a theory of *O*'s perceptual mechanism, a theory designed to explain how *O* visually estimates length; and that Psychophysical Law (5), which is the mathematical completion of this theory, depends on the mak scale. The argument for this theory of *O*'s perceptual mechanism is that failure to accept it will leave unexplained *O*'s ability to make reliable estimates of physical length. This argument is far from conclusive.

In the first place, it is doubtful, as we have argued, that (Ca) and (Cb) are verifiable. If they are not, then the theory in which they figure is not scientifically respectable. Secondly, alternative theories of the same type with equal explanatory power can be obtained by replacing (Cb) with some other assumption, by assuming that physical estimates of 1:2 are based on psychological

estimates of 1:3, or 1:4, or 2:3, etc. Thirdly, it has yet to be shown that any of the alternative theories just indicated are required to explain O's ability to make estimates of relative physical length. Why must we suppose that O bases his estimates of physical ratios on estimates of psychological ratios? Why isn't the following explanation sufficient? The O's eyes are normal, and he has learned how rods look when they stand in the ratio 1:2; hence, when he looks at rods he can say with some accuracy whether or not they stand in this ratio. That this explanation implies no perceptual mechanism like that in (Ca)-(Cb)-(Cl) is not enough to reject it.

It may be well to include a warning against attempting an epistemological justification of ratio scaling. The essential contention of the introspectionist theory of O's perceptual mechanism is that O bases his (always indirect) estimates of physical entities on his (always direct) estimates of psychological entities. This sort of perceptual theory is often subsumed under or associated with the philosophical, epistemological theory that knowers base their (always indirect) knowledge of the external environment on their (always direct) knowledge of private, internal processes, these last being causal results or at least reflections of the environment. This epistemological theory is clearly dualistic—since it posits parallel psychological and physical realms, and introspectionistic—since O is believed to know the internal realm by some process of inner perception, that is, by "introspection." The theory seems both to establish our knowledge of the external world and to explain how we obtain it, and it has, as a consequence, attracted psychologists and philosophers for centuries.

But there are problems, the most important of which is an instance of the sceptical problem of solipsism. If O directly perceives only internal entities, then how can he know that his inferences concerning the existence and character (physical length, for instance) of external entities are accurate? Some theorists may suppose that this problem is solved by the discovery of such laws as (5), that if O learns from E the precise relation between his internal entities (Ψs) and external entities (Φs), then he can make reliable inferences from the ones to the others. (Gibson [1950, pp. 186–187] seems to make some such supposition. He is properly taken to task for it by Price [1953, p. 410].) But the sceptical problem which confronts O also confronts E. The E also directly perceives only internal entities; consequently, his inferences concerning the physical length of the experimental rods are equally problematic. Since E must determine the physical lengths of the experimental rods in order to establish Psychophysical Law (5), and since E's physical length inferences are just as problematic as O's, O cannot rely on the validity of (5) to support his own inferences concerning physical length. The blind cannot be led by the blind. (This difficulty is related to the "psychologist's circle" which Boring [1931], Bergmann & Spence [1944, pp. 2–5], and others have discussed.) It is no reply to say that E *measures* the rods—with a meter stick, say—and thus insures the accuracy of his determinations of their length. For E's measurements are accurate only if he correctly infers the physical lengths of meter-stick segments from their psychological lengths. And how can he know that these inferences are correct?

If, as it appears, the solipsist problem arises in an introspectionist but not in a behaviorist interpretation of psychophysical measurement, then this is clearly an argument in favor of the latter.

The Behaviorist Interpretation

The Principle of Correspondence

It is often said that scales like the one constructed in Experiment M are "scales of observer response." This statement is either imprecise or false. For it suggests that psychological magnitudes are literally magnitudes of O's responses: perhaps the loudness of his verbal reports, or their pitch, or the time required to make them, etc. Obviously none of these suggestions is acceptable. There is no reason to suppose that the loudness, pitch, or time of O's verbal reports has any interesting or systematic relation either to the length of the rods or to estimates of their length. The behaviorist must not *identify* psychological length with some

magnitude of O's responses; rather, he must *define* psychological length in terms of O's responses in some useful manner. Many of Stevens' (1959a, p. 52) remarks suggest the erroneous identification, for example: "brightness . . . is the name for a response of a human organism to an external configuration of the environment."

Since every magnitude—psychological or otherwise—is defined by the relational terms "greater than," "equal to," and "less than," we must, in defining psychological length, provide a rule for the application of at least these three terms. Such a rule is contained in Definition (i): If O estimates that Φ_1 is (physically) greater than (equal to, less than) Φ_2, then Ψ_1 is (psychologically) greater than (equal to, less than) Ψ_2. This statement defines psychological length as an intensive magnitude, but fails to give it all the features of the extensive magnitude scaled in Experiment M. To generate an extensive magnitude a definition of psychological ratios is also required. Hence we need also Definition (ii): If O estimates that Φ_1 and Φ_2 stand in the (physical) ratio $n{:}m$, then Ψ_1 and Ψ_2 stand in the (psychological) ratio $n{:}m$. (i) and (ii) can now be used, together with an arbitrarily chosen unit of psychological length, to construct the ratio scale of psychological length in Figure 2.

Definition (ii) is, of course, the principle of correspondence encountered earlier. But by making the principle a definition the behaviorist denies that it requires a supporting argument such as (Ca)-(Cb)-(Cl), a theory of O's perceptual mechanism. We argued that (Ca) is dubious and apparently unverifiable, and that (Cb) is dubious and potentially circular. We argued that there are alternative theories to the one embodied in (Ca)-(Cb)-(Cl). The behaviorist is unaffected by these objections, since he subscribes neither to (Ca) and (Cb), nor to any theory of O's perceptual mechanism which posits internal estimates as the basis for external estimates. He simply begins with the fact that O's responses to rod length are reliable, and then goes on to use the mak scale as a means for expressing mathematically the relation between those estimates and actual rod length.

Though immune to some objections, the behaviorist interpretation of the principle of correspondence may be open to others. If the principle is merely a definition, then why should we accept it? The reply cannot be that O's estimates of Φ ratios are based on his estimates of Ψ ratios and his estimates of Ψ ratios are accurate. This would be to fall back on the introspectionist view. Nor can the behaviorist maintain that (ii) is a natural definition, that is to say, a definition of ordinary terms, like "A vixen is a female fox." Psychological length is a technical notion which the behaviorist psychophysicist introduces for extraordinary purposes. (ii) must, therefore, be regarded as a stipulative definition, which can be supported only by appealing to its scientific usefulness. Whether it is in fact useful will be discussed in the sequel.

(i) and (ii) are examples of what are sometimes called "operational definitions" of a psychological magnitude. Most behaviorist psychophysicists call for such definitions, assure the reader that they can be provided, and then fail to provide them. Stevens (1935) illustrates this sort of malpractice. Goude (1962, pp. 28–29) may be an exception. In failing to provide explicit statements of the definitions, the psychophysicist risks overlooking several troublesome but extremely important questions concerning the behaviorist interpretation. One such question is: Are there as many types of psychological magnitude as there are types of estimate? Are psychological magnitudes constructed from different types of estimate incomparable?

Suppose O estimates in a halving experiment for length that

$$\Phi_1 = \tfrac{1}{2}\Phi_2 \quad \text{and} \quad \Phi_2 = \tfrac{1}{2}\Phi_3\,, \tag{7}$$

and in an experiment requiring length estimates of one-third, he says that

$$\Phi_1 = \tfrac{1}{3}\Phi_3\,. \tag{8}$$

Together with Definition (ii), (7) entails that

$$\Psi_1 = \tfrac{1}{2}\Psi_2 \quad \text{and} \quad \Psi_2 = \tfrac{1}{2}\Psi_3\,. \tag{9}$$

Together with certain elementary rules of algebra, (9) entails that

$$\Psi_1 = \tfrac{1}{4}\Psi_3\,. \tag{10}$$

On the other hand, (8) together with Definition (ii) entails that

(11) $\Psi_1 = \frac{1}{3}\Psi_3$.

Apparently (10) and (11) are inconsistent; consequently, the estimates in (7) and (8) apparently conflict. The behaviorist may try to remove the inconsistency by maintaining that (*a*) Definition (ii) is not applicable to all *O*'s fractionations, (*b*) the elementary laws of algebra do not apply to psychological magnitudes, (*c*) *O* is mistaken in some of his estimates of psychological length, or (*d*) (10) and (11) describe two different psychological magnitudes. (*a*) is an unacceptable solution, since it undermines the definition which makes the construction of ratio scales of psychological length possible. (*b*) is also unacceptable, since it implies that the numerals assigned to psychological entities in experiments like M do not represent ratios. (*c*) is completely out of the question, since it carries the implication that psychological entities are observable, and thus plunges us back into an introspectionist framework. Only solution (*d*) remains.

The estimates on which Figure 1 is based do not lead to the sort of inconsistency illustrated above. But there is no assurance that all or even most actual experiments will be like M in this respect. It is always possible—indeed, it is likely—that *O* will make conflicting estimates. What the behaviorist is forced to say about conflicting estimates shows that, even in those experiments which contain no conflicts, different fractional estimates create different psychological magnitudes. If this is so then it is wrong to suppose that in Figure 2 Points A-E and Points a-c lie on the same psychological continuum. The two sets of points do lie along the same *line*. But this is nothing more than an artifact of particular estimates made by a particular observer. The same observer at another time, or another observer, may easily produce conflicting estimates. Then the two sets of points will not even lie on the same line.

Where we compare ratio estimates with interval estimates the "operationist" features of Definition (ii) become even more obvious. Suppose *O* halves three rods as in (7), and then estimates length intervals for the same rods by saying that

(12) $\Phi_3 - \Phi_2 = \Phi_2 - \Phi_1$

Together with a principle of correspondence for interval estimates, (12) entails that

(13) $\Psi_3 - \Psi_2 = \Psi_2 - \Psi_1$.

Apparently (9) and (13) are inconsistent; consequently, the estimates in (7) and (12) apparently conflict. The behaviorist may try to remove the inconsistency with solutions analogous to (*a*), (*b*), or (*c*). Those solutions will be unacceptable for similar reasons. The conflict can be acceptably removed only by maintaining that (9) and (13) describe different psychological magnitudes, different psychological lengths.

In brief, if we hold that principle of correspondence (Cl) and other similar principles are stipulative definitions, then we seem forced to admit that there are as many types of psychological length as there are types of estimate of physical length, that there is no single continuum called "psychological length." If this is so, then no scale of psychological length can conflict with any other. A ratio scale constructed from m/n estimates cannot conflict with a ratio scale constructed from n/o or p/r estimates. Hence, Campbell (in Ferguson, et al., 1939–40, p. 338) is mistaken in his criticism of ratio-scale measurement. He argues that if a loudness scale is constructed from estimates of one-half, then "x sone will *not* be estimated as a tenth of $10x$ sone." What then, he wonders, is the advantage of a figure like our Figure 2 over one like Figure 1? "Why do not psychologists accept the natural and obvious conclusion that subjective measurements of loudness in numerical terms (like those of length or weight or brightness) are mutually inconsistent and cannot be the basis of measurement?" The behaviorist ought to reply that one-half scales and one-tenth scales are scales of different psychological loudnesses, and cannot be inconsistent.

More obviously, a ratio scale constructed from fractionation estimates cannot conflict with a partition scale constructed, say, from equisection estimates; and neither of these

can conflict with a jnd scale. It may be supposed that since none of the three curves in Figure 2 coincides with any of the others, only one can represent a "valid" scale of psychological length. But this is true only if the ratio scale, S_R, the partition scale, S_P, and the jnd scale, S_J, are all scales of one and the same psychological magnitude, Ψ_A. Now according to the behaviorist, S_R, S_P, and S_J are scales of different psychological magnitudes, Ψ_A, Ψ_B, and Ψ_C, respectively. Hence, the three scales do not compete with one another, and the fact that they do not coincide raises no question about their validity.

Discussing the lack of correspondence between ratio and partition scales, Stevens (1960c, pp. 52–53) says: "... observers are so constituted that they are unable to partition a prothetic continuum without a systematic bias." He tries to explain this fact by suggesting that the observer's sensitivity is not uniform on the scale, being greater in the lower ranges. To say that partition estimates exhibit a systematic bias implies that the partition scale and the scale with which it is being compared are scales of a single psychological continuum. More precisely, the implication is that S_P and S_R are scales of Ψ_A, that S_P and S_R do not coincide, and that S_R is the "true" scale. The behaviorist reaction is that S_P and S_R are scales of different psychological magnitudes and therefore do not compete. The one is as "true" a scale as the other; so the estimates producing the one are no more biased than the estimates producing the other.

Stevens sometimes takes a different tack. At one point (1959c, p. 996) he writes that the question of scale validity is "a matter of opinion," that "a judgement about validity always reduces ultimately to a value judgement," and that "in the long run ... it is the scientific community that will decide the issue." There is no issue for the behaviorist, since different types of psychological scale are scales of different psychological magnitudes and do not compete for validity. At other points, Stevens adopts this "operationist" point of view.

> Since the three kinds of scales are nonlinearly related on prothetic continua, it seems clear that they must measure different things. Each is probably a valid scale of something [1959c, p. 998].

Speaking of the three different types of scale for subjective finger span, he says:

> Obviously, three different aspects of finger span are being measured by these three functions. Although a certain amount of argument has revolved around the question of which of these functions is the "true" scale, it should be apparent that all three are true scales of something or other [Stevens & Stone, 1959, p. 94].

It is impossible to locate Stevens' view precisely, since he constantly shifts back and forth between a behaviorist and some other way of treating the question of scale validity.

The Nature of Psychological Entities

According to the behaviorist, the Ψs defined in (i) and (ii) are not observed by *E*, *O*, or anyone else. This harmonizes the interpretation with the phenomenological facts of Experiment M, which are as follows. The *O* sees rods and walls, feels chairs and tables, hears voices, and so on. But he does not observe, through some mysterious faculty of perception, psychological entities of which psychological length is a magnitude. The *O* observes by sight the experimental *rods*, and he estimates their relative *physical* length. The *E* measures the *physical* length of the *rods*, records *O*'s estimates, and constructs a psychological magnitude in accordance with Definitions (i) and (ii). Neither *O* nor *E* observes psychological entities or magnitudes.

This feature of the view has an important bearing on the question of observer accuracy. Psychophysicists often say that experiments like M presuppose the ability of the observer to make accurate estimates of psychological magnitudes. Thus Stevens (1951, pp. 40–41) says:

> [In fractionation procedures] we make an assumption that calls for scrutiny. We postulate, among other things, that the subject knows what a given numerical ratio is and that he can make a valid judgement of the numerical relation between two values of a psychological attribute.

Such a statement can be understood only in the context of an introspectionist interpretation of scaling experiments. As a previous section has shown, the introspectionist bases

principle of correspondence (Cl) on the assumption, (Cb), that *O*'s estimates of psychological magnitude are accurate. But the behaviorist maintains that *O* does not make estimates of psychological magnitudes, that *O*'s estimates are of physical magnitudes like length, weight, etc. Now it is clear that Experiment M does not presuppose *O*'s ability to make accurate estimates of physical length. Indeed, *O*'s length estimates are inaccurate, which is typical of experiments of this sort. And *E* will often give *O* explicit instructions to provide naïve estimates, to make no attempt at being accurate. Thus Garner tells his subjects: "Remember to try to assign numbers according to how loud the tones appear to *you*. We are interested in how loud tones *seem* to be to you, not in some kind of accuracy" (quoted by Stevens [1956, p. 17]). In sum, on the behaviorist view, *O* makes no estimates of psychological magnitudes, and the estimates which he does make—estimates of physical magnitudes—are not required or presupposed to be accurate.

There are also important implications for a related question, that of the so-called "stimulus error." As the term was originally introduced by Titchener (1905, p. xxvi), to commit the stimulus error is (*a*) "to confuse sensations with their stimuli," "to read the character of the stimuli into the 'sensations'." As Boring (1921, p. 451), who reviewed the history of the notion, put it: "We commit the stimulus-error if we base our psychological reports upon objects rather than upon the mental material itself, or if, in the psychophysical experiment, we make judgements of the stimulus and not judgements of sensation". However, the term has been used in a different, or at least extended way. Stevens (1959c, pp. 1002–1003), in criticizing the physical correlate theory, characterizes it as holding that "*all* quantitative estimates of sensory magnitude are really based on some form of 'stimulus error'." The theory under attack was devised by Warren (1958) and Warren, Sersen, and Pores (1958), who put it to experimental test with loudness. (See Warren & Warren [1963, pp. 804–808] for a further discussion of the theory.) They maintain that, knowing little or nothing about the magnitudes of sound waves, typical observers base their estimates of loudness on the (estimated) distance of the sound source. Now, to commit this error is not to confuse sensations with their stimuli, but rather (*b*) to confuse one stimulus with another. Thus we have a different meaning of the term "stimulus error."

The behaviorist may with perfect consistency attack the physical correlate theory. He may regard it as an error to confuse one stimulus with another, and he may attempt to show by experiment that observers do not commit the error. (He may also, again with perfect consistency, attempt to show that observers *do* commit the error when estimating loudness, brightness, etc. But what, in Sense [*b*], would the stimulus error be for length, weight, etc?) The behaviorist may, therefore, attack the stimulus error in Sense (*b*). But he may not attack it in Sense (*a*). For on his view, typical observers do not commit any error when they make estimates of the stimulus in psychophysical experiments. That is precisely what they are asked by *E* to do. Indeed, the behaviorist cannot even admit the *possibility* of stimulus error in Sense (*a*). To commit the error in that sense is to be aware of the stimulus, to be aware of the sensation caused, and to "read" the former into the latter. But the behaviorist holds that observers are not aware of any sensations caused by rods, weights, etc. in psychophysical experiments dealing with these stimuli. The behaviorist experimenter may desire that his observers estimate length, weight, etc. "as they see it," "as they feel it," etc., and he may give them instructions to that effect. But these will, nonetheless, be instructions to estimate the *stimulus*.

One decided advantage in the behaviorist interpretation is its removal of private entities from the conceptual structure of Experiment M. By definition, a public entity can be observed by any normal observer, a private entity by only one observer. Hence, the distinction between public and private entities applies only to observable entities, like rods and afterimages, and not to the unobservable Ψs of the behaviorist interpretation. It follows that we must not say that Ψs are public. But, when we recog-

nize the parallel mistake in saying that they are private, there is no longer any inclination to regard them as directly inaccessible to everyone but *O*, or to think of *O* as an intermediary between the investigating scientist and a realm of private data. The dualism between public and private data entirely collapses, and, as a result, several disturbing theoretical and philosophical problems simply vanish. Psychological entities are indeed unobservable; nevertheless, since they are defined in terms of publicly observable rod presentations and observer reports, there is no problem about the general availability of psychological data or "objectivity" of results based on these.

According to the behaviorist, Ψs are, to use a current phrase, theoretical constructs. This view is suggested vaguely by Garner (1954, p. 88) and incompletely by Stevens (1960a, p. 27). That is, they are *theoretical* entities *constructed* by *E* by means of Definitions (i) and (ii) for scientific uses. An analogy will help to explain their nature. The *O* is asked in a two-part experiment to estimate the dollar value of automobiles. In each trial of the first part, *E* presents *O* with a standard auto and asks him to select a comparison auto one-fourth as expensive. In each trial of the second part *O* is asked to select comparison autos which are one-half as expensive as the standards. A plot of the data thus obtained produces lines identical to those in Figure 1, except that numerals along the axes represent auto values in thousands of dollars. The *E* now invokes a principle of correspondence, which says: If *O* estimates that Φ_1 and Φ_2 stand in the ratio 1:2, then Ψ_1 and Ψ_2 stand in the ratio 1:2, where Φs represent the real value and Ψs the estimated value of automobiles. The *E* stipulates that one unit of estimated value equals 1,000 units (dollars) of real value, and, using the principle of correspondence above, constructs a scale of estimated value, plotting estimated against actual value by the same method used to obtain Figure 2.

It is clearly a mistake to attempt to identify the Ψs in our analogy with some group of privately observable entities, of which estimated value is a magnitude. This attempt will lead to theories like the following. When *O* estimates the value of a $5,800 automobile he has thought, Ψ_1; when he estimates the value of a $10,000 automobile he has another thought, Ψ_2. Since Ψ_1 is (psychologically) half as great as Ψ_2, *O* estimates that the values of the two automobiles stand in the ratio 1:2. This theory is specious, firstly, because no meaning can be given to the assertion that one thought is less great or greater than another; and, secondly, because it is an obvious fact that *O* makes estimates, not about his thoughts, but about his automobiles. Similarly, it is an introspectionist mistake to try to identify the Ψs of the mak scale with privately observable sensations. Such identification commits us to the unclear and possibly meaningless view that sensations can be greater or less than one another; and it implies, falsely, that *O* makes estimates of sensations.

It is also a mistake to ask for examples of the entities of which estimated value is a magnitude. Estimated value cannot be construed as a magnitude of *O*'s utterances (their loudness, time, etc.), nor of automobiles, nor as a magnitude of any other observable group of entities. Analogously, it is a mistake to ask for examples of the entities of which psychological length is a magnitude. Behind this request lies the introspectionist belief that Ψs can be ostensively defined and are thus observable entities, examples of which can literally be pointed to by someone. One completely understands the nature of psychological entities and their magnitudes when he understands Definitions (i) and (ii). To ask for examples, after having carefully studied the definitions, betrays misunderstanding.

In view of the above considerations, it is tempting to say that on the behaviorist view psychological entities are fictions and do not exist. Although true in one sense of "exist" (the sense in which observables exist), this is apt to be misleading, to suggest that the psychophysicist is trying to measure nothing. It is more helpful to point out that the concept of psychological magnitude is the product of a *façon de parler*. Consider again our analogy. When we say, (*a*): "The estimated value of the one automobile is half the estimated value of the other," we do not wish to create the impression that automo-

biles have two sorts of value, estimated and actual value, in the way that they have ballast value on a ship as well as financial value on the market. (*a*) is just another way of saying, (*b*): "*O* estimates that the value of the one automobile is half the value of the other." (*a*) means nothing more than what is meant by (*b*), and (*b*) does not imply that estimated and actual value are two sorts of value possessed by automobiles. Only money value is involved, although an economist or someone else may be interested in O's estimates of this money value. (Other examples: "alleged" and "confirmed" age are not two sorts of age, nor "probable" and "determined" area two sorts of area, nor "apparent" and "actual" cause two sorts of cause.) Analogously, when we say, (*c*) "The estimated length of the one rod is half the estimated length of the other," we do not imply that estimated length and actual length are two sorts of length possessed by rods. (*c*) is just another way of saying, (*d*) "*O* estimates that the length of the one rod is half the length of the other." And (*d*) carries no implication of two sorts of length. Rods possess one sort of length: physical length, if you will. But the psychophysicist is interested in the estimates which observers make of the physical length of rods. The *E* may express facts about such estimates by speaking of "estimated length" as opposed to "actual (real) length." But it is clear that such talk is merely a *façon de parler* which *E* finds it convenient to use in describing the behavior of his observers.

It is of interest to bring our analogy to bear on Stevens' (1959a, pp. 50ff.) suggestion that utility can be measured by the scaling procedures used for perceptual magnitudes. He describes an experiment by Galanter in which *O* is given instructions like the following: "Suppose I were to tell you that I am going to give you $10. That would make you happy, would it not? All right, now think this over carefully. How much would I have to give you to make you twice as happy?" The *O*'s responses are used to construct a scale of "utility," or "subjective value," measured in "utiles." (Presumably the method is like that used in deriving Figure 2 from Figure 1.) This experiment differs slightly from the one in our analogy. There *O* was estimating the dollar value of a class of *goods*. Galanter's observer makes estimates of *dollars*. But he does not estimate the dollar value of dollars. Such estimates would be nonsensical, unless *O* were estimating the new dollar value of dollars in terms of their old dollar value, which he is not. Rather, *O* estimates the capacity of money to make him happy. And the utility, or subjective value, of money is defined in terms of these estimates. Hence, to say, (*e*) "$18 has twice as much utility (subjective value) as $10 for *O*" is just to say, (*f*) "*O* estimates that $18 has twice the capacity of $10 to make him happy." Employed in this way, the concept of utility is the product of a *façon de parler*. The same analysis applies to psychological length, as the behaviorist employs that concept. The point of the *façon de parler* is, of course, to make the phenomena spoken of amenable to measurement.

The Purpose of Ratio Scaling

In a previous section we noted that the behaviorist can argue the acceptance of principle of correspondence (Cl) only on grounds of scientific usefulness. Now obviously the reason for accepting (Cl) is that it makes the construction of the mak scale possible. But what reason is there for constructing the scale? Well, its construction makes the formulation and verification of Psychophysical Law (5) possible. But what reason is there for formulating and verifying (5)? The behaviorist cannot answer, in the manner of the introspectionist, that (5) is a law relating two variables, Ψ and Φ, defined independently of one another, and that the discovery of such laws needs no justification. For Ψ is defined in (i) and (ii) in terms of *O*'s Φ estimates. Nor can the behaviorist plead that (5) explains how *O* makes reliable estimates of physical length. (5) could provide such an explanation only in conjunction with a theory of *O*'s perceptual mechanism like (Ca)-(Cb)-(Cl). But the behaviorist does not link (5) to any such theory. Consequently, he appears reduced to saying this: (5) does not tell us *how* *O* estimates physical length; rather it is a mathematical description of *what* *O* estimates physical length to be. (5) is (to employ some current terminology) not

an *explanation;* rather it is a mathematical *description* of O's estimates of physical length. The reason for accepting (Cl) is that it makes this mathematical description possible.

The behaviorist may complain that (5) is more than a mere description of O's length estimates, since it enables us to predict new data, to predict ratio estimates which were not made by O in obtaining the data on which Figure 1 is based. By using (5) we can predict for any Φ_s which Φ_c O would estimate to be one-half, one-fourth, etc. Now isn't predictive power what distinguishes an explanatory law from a merely descriptive one? Isn't any explanatory law, (5) or any other, simply one with predictive power? This reply suggests the view that (*a*) psychological explanation consists in the discovery of empirical laws which describe and predict the behavior of organisms. The opposing view is that (*b*) psychological explanation consists in discovering the mechanisms, mental or physiological, which underlie the behavior of organisms.

It is not necessary for us to ask which of these two views is correct, nor to ask which is behaviorist and which is introspectionist. We will remark only that there seems to be no reason why a behaviorist cannot accept explanations in terms of *some* mechanisms, i.e., those which are physiological and do not posit internal estimates as the basis for external estimates. The solution to these difficult questions does not affect the difference between the introspectionist and behaviorist justifications of Psychophysical Law (5). If the behaviorist adopts (*b*), then he is required to say that (5) is descriptive and not explanatory, since it is not, on his view, a theory or part of a theory of O's perceptual mechanisms. If he adopts (*a*), then he may claim that (5) is in one sense an explanatory law, since it has predictive power. But he must deny that it is explanatory in the other sense. For he denies that O perceives and makes estimates concerning external, physical entities by perceiving and making estimates concerning internal, psychological entities. Whereas, such an hypothesis is essential for (5) to be a theory or part of a theory of O's perceptual mechanism.

If the introspectionist justification cannot be adopted by the behaviorist, then what, on the latter's view, is the purpose of ratio scaling? That Psychophysical Law (5) can be used in describing and predicting O's fractionation estimates may appear to be sufficient reason for employing any scaling procedure which enables us to formulate and verify the law. But is this so? (3) can be used in describing and predicting O's one-fourth estimates, (4) in describing and predicting O's one-half estimates. Other laws of the same type can be used in describing and predicting other fractionation estimates. None of these laws presupposes the concept of psychological length. It is obviously useful to construct the lines and discover the equations of Figure 1. But what additional purpose is served by going on to construct the line and formulate the equation in Figure 2, a construction which presupposes the concept of psychological magnitude? More generally, the question is this: Is there any legitimate purpose in constructing ratio scales of psychological magnitude which cannot be equally well achieved without the concept of psychological magnitude? If the answer is negative, then the concept of psychological magnitude may and should (for reasons to be given later) be dispensed with. We will attempt to discredit three arguments which conclude that the concept is indispensable.

First argument. One argument insists that without the concept of psychological magnitude we would not be able to formulate the *general* psychophysical laws which have been advanced in the past few decades. (3), (4), and equations for other estimation-ratios can be obtained from (5), and (5) can in turn be obtained from them. This can be accomplished by using the equation

(14) $$a = b^n,$$

or, logarithmically,

(15) $$n = \log a/\log b,$$

where a is the fraction (one-fourth, one-half, etc.) O is required to estimate, $b = \Phi_c/\Phi_s$, and n is the exponent in an equation like (5). Thus it is possible to perform the useful operation of predicting the results of new fractionation experiments. (Ekman [1958, p. 288]

and Treisman [1964b, p. 387] both suggest an algebraic method which is essentially the same as that offered here.) In view of the algebraic relations above, (5) may be considered as a kind of general law, of which (3), (4), and certain other fractionation equations are, in a sense, instances. Now it may be supposed that such general laws must take the form of (5), that is, must relate some psychological magnitude to a stimulus magnitude.

This claim appears to be false. Given (14), we can write

$$(16) \qquad b = \sqrt[n]{a},$$

and can then substitute into the equation, $\Phi_c = b\Phi_s$, to obtain:

$$(17) \qquad \Phi_c = \sqrt[n]{a}\,\Phi_s\,.$$

Where n has the value determined in Experiment M, (8) becomes

$$(18) \qquad \Phi_c = \sqrt[2.16]{a}\,\Phi_s\,.$$

(3), (4), and all the other fractionation equations which can be obtained from (5) by the method in the previous paragraph can also be obtained from (18) by the same method. And from every fractionation equation from which we can obtain (5) by the algebraic method in the previous paragraph we can also obtain (18) by the same method. Thus (18) is as general, and as useful in prediction, as is (5); and yet it does not employ the concept of psychological length.

Since the above argument assumes that all fractionation equations are linear, its conclusion is not a general one. The general claim is as follows: Given any equation, $\Psi = F(\Phi)$, from which we can obtain and which can be obtained from any of the fractionation equations, $\Phi_c = f_1(\Phi_s)$, $\Phi_c = f_2(\Phi_s)$, $\cdots$ $\Phi_c = f_k(\Phi_s)$, by the method described above; there is another equation, $\Phi_c = G(\Phi_s)$, which we can obtain from and which can be obtained from any of the same fractionation equations by the same method. If this claim is true, then the most general psychophysical laws can be formulated without employing the concept of psychological magnitude. The G equation is of an entirely different type than the F equation, since it contains only physical variables. But it is, nonetheless, a psychophysical equation, an equation which can be used to describe and predict O's ratio estimates of physical length.

Second argument. This argument maintains that without the concept of psychological length we would be unable to make the comparisons made in Figure 2 between ratio scales, partition scales, and jnd scales. Thus stated the argument is trivial and question begging. Naturally, if we do not construct a scale, S_R, of psychological length, then we cannot compare it with scales S_P and S_J, since there is nothing with which to make the comparison. The question is: Why should we construct any scales of psychological length? There may indeed be good reasons for constructing psychological scales of some kind. But is there any good reason for constructing *scales of psychological magnitude*? Why even introduce the concept of psychological magnitude? The answer will be that if we do not we cannot compare ratio estimates with interval estimates, nor either of these with jnd estimates.

This argument is unconvincing, since there seem to be other ways of comparing the various types of estimate. Let us illustrate an alternative method by comparing jnd estimates with ratio estimates of one-half. The upper line in Figure 2 is the result of plotting number of jnd's against rod length. By using this line we can plot in Figure 1 the length of the stimulus associated with n jnd's (ordinate) against the length of the stimulus associated with $2n$ jnd's (abscissa). For example, Figure 2 shows that the stimuli associated with 20 and 40 jnd's are 5 and 11 cm., respectively. Accordingly, we place a point at 5 on the y axis and 11 on the x axis of Figure 1. Arbitrarily selected points obtained in this manner are connected by the lower of the two dash lines in Figure 1. That the dash line does not coincide with Line (4) presumably shows that O does not estimate length ratios of one-half on the basis of length jnd's. (If coincidence had been the result, would this have shown that O *does* estimate length ratios of one-half on the basis of length jnd's?) Similarly, we can compare jnd estimates with ratio estimates of one-fourth by plotting in Figure 1 the length of the stimulus associated with n jnd's

against the length of the stimulus associated with $4n$ jnd's.

We can use this same method to compare partition estimates with ratio estimates, so long as the partition interval is relatively small and the smallest stimulus estimated is near physical zero. The numerals on the inside left-hand ordinate of Figure 2 can be taken to represent merely the *number* of apparently equal intervals associated with a given stimulus, just as the numerals on the outside right-hand ordinate represent the number of jnd intervals associated with a given stimulus. Then we may plot in Figure 1 the length of the stimulus associated with n apparently equal intervals against the length of the stimulus associated with $2n$ apparently-equal intervals. If we do this we obtain the upper of the two dash lines in Figure 1. That this dash line does not coincide with Line (4) presumably shows that O does not estimate length ratios of one-half on the basis of equal-appearing intervals.

It is important to understand that in our method of comparison, the ordinate numerals of the category scale and jnd scale represent, not the *size* of intervals along a psychological continuum, but merely the *number* of intervals associated with various stimuli. Our method makes the harmless and probably trivial assumption that jnd's are similar in some respect, and that apparently equal intervals are similar in some respect; but it does not assume that jnd intervals, or apparently equal intervals, are equal on some psychological continuum. (Similarly, to count off the number of octaves in the musical scale assumes that octaves are similar in some respect, but not that they are equal on a psychological continuum.) If the ordinate numerals represent equal intervals on a psychological continuum, then the following principle of correspondence is presupposed: If O estimates that $\Phi_1 - \Phi_2 = \Phi_2 - \Phi_3$ then $\Psi_1 - \Psi_2 = \Psi_2 - \Psi_3$. But our method presupposes no such principle. We can speak of the number of similar intervals without employing the concept of a psychological entity or psychological magnitude. Our method thus makes it possible to compare O's jnd estimates and partition estimates with ratio estimates without measuring,

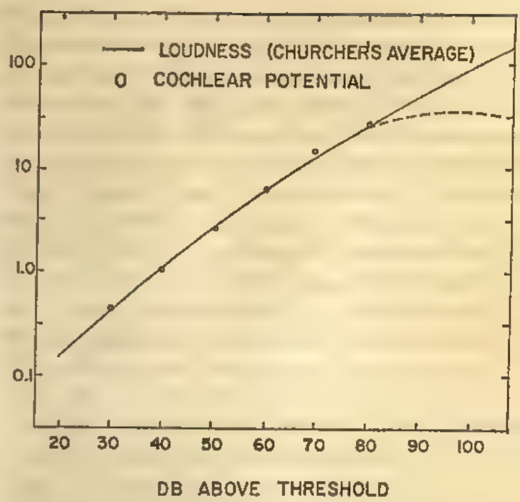

FIG. 3. The loudness function and the size of the cochlear potential. The solid curve represents the loudness function arrive at by Churcher from tone-fractionation experiments. The circles represent the averaged results of measurements of the size of the cochlear potentials at the round windows of six guinea pigs as a function of the intensity of the stimulus. (Taken from Stevens & Davis, 1936, p. 5.)

scaling, representing, mentioning, or presupposing a psychological magnitude called psychological length. This is not to suggest that it is the only method with this feature. Avoiding the concept of psychological magnitude seems to depend only on the ingenuity of the theorist or experimenter in representing the data of psychological experiments.

Third argument. This argument contends that without the concept of psychological magnitude we would not be able to compare psychophysical functions with physiological functions. Figure 3, taken from Stevens and Davis (1936, p. 5) shows that when cochlear potential (multiplied by a constant) and psychological loudness are plotted against sound-wave intensity, the curves thus obtained show a high degree of correspondence. (That the loudness function in this figure is no longer accepted is irrelevant to our argument.) That is to say, where $\Psi = f_1(\Phi_1)$ is the psychophysical function and $\theta = f_2(\Phi)$ the physiological function (θ designates the physiological magnitude), f_1 is linearly related to f_2. Consequently, the authors suggest that "as a

first approximation, the form of the loudness function is imposed by the behavior of the cochlear mechanism;" but they point out that "identification of the loudness function with the recorded potential must [in view of divergences] be made with reservations" (p. 6). Now it may be said that such interesting suggestions could not be made without employing the concept of psychological magnitude.

Thus stated, this argument begs the question. Let Ψ and θ be a psychological and a physiological magnitude, respectively. It is true that we cannot compare a psychophysical function, $\Psi = f_1(\Phi)$, with a physiologicophysical function, $\theta = f_2(\Phi)$, without employing the concept of psychological magnitude. But this observation is trivially true, since, by hypothesis, one of the functions to be compared contains Ψ as a variable. If we are to avoid begging the question, we must ask whether the results of psychophysical experiments can be compared with the results of physiologicophysical experiments without employing the concept of psychological magnitude, whether such comparisons can be made without comparing Ψ-Φ functions with θ-Φ functions. The answer is, apparently, that it is possible.

Let θ_c be the physiological process (cochlear potential, area of retinal stimulation, etc.) associated with the physical comparison stimulus, Φ_c (sound-wave intensity, rod length, etc.). And let θ_s be the physiological process associated with the standard stimulus Φ_s. In addition to fractionation functions of the form, $\Phi_c = f_3(\Phi_s)$, we can also formulate and verify associated functions of the form $\theta_c = f_4(\theta_s)$. If $f_3 = f_4$, or if f_3 and f_4 are linearly related, then we may wish to conclude that θs are the physiological processes underlying or determining ratio estimates of Φs. The comparison and discovery of functions of this type can be accomplished without measuring, scaling, representing, mentioning, or presupposing psychological magnitudes. (The measurement of physical and physiological magnitudes *is*, of course, required; but there is no problem here.) Since Φ-Φ and θ-θ functions appear to be as useful as Ψ-Φ and θ-Φ functions in comparing the results of psychophysical and physiological experiments, and since the former functions do not employ the concept of psychological magnitude, why should we retain the concept? No claim is made here that the comparison method suggested is the only possible one which avoids the concept of psychological magnitude. The point is rather that avoiding the concept seems to depend solely on the ingenuity of the experimenter or theorist in representing the results of relevant experiments.

An additional point of considerable force against each of the foregoing arguments for psychological magnitude is that principle of correspondence (Cl) is, in an important sense, arbitrary. Ekman & Sjöberg (1965, p. 452) may have some point like this in mind when they say: "According to a strictly behaviorist view of perception . . . the [psychological] scale is an arbitrary and possibly trivial transformation of response data . . . " Instead of (Cl), why not assume that when *O* estimates that Φ_1 and Φ_2 stand in the ratio 1:2, Ψ_1 and Ψ_2 stand in some other ratio, say 4:7? If we make this assumption, then (5) can no longer be derived from (4), so that (3) and (4) can no longer be regarded as instances of the single general Psychophysical Law (5). Furthermore, if we change the psychological unit (to 1 mak = 4 cm.), a Ψ-Φ curve can be derived by employing the new principle which is in closer correspondence with the partition curve in Figure 2. Then it becomes easier to argue that ratio and partition scales are equally "valid" scales of psychological length. Finally, the Ψ-Φ curve derived from the new principle may be in better, or worse, correspondence with any θ-Φ curve which has been discovered. If agreement is worse then it may no longer be possible to argue that the θ-Φ curve describes the physiological mechanism underlying *O*'s ratio estimates. The solid line in Figure 3 was derived by making use of an assumption like (Cl) for loudness-frequency, and also the assumption that a tone introduced into one ear sounds half as loud as the same tone introduced into both ears. If these assumptions are replaced by others then it is no longer possible to argue that "the form of the loudness function is

imposed by the behavior of the cochlear mechanism."

By assuming still other principles of correspondence, Experiment M can be made to verify a psychophysical log law instead of a psychophysical power law. Each of the alternative principles of correspondence so far considered assume that equal physical ratios are accompanied by equal psychological ratios. That is to say, they are all instances of principle (C): If O estimates that $\Phi_1/\Phi_2 = \Phi_2/\Phi_3$ then $\Psi_1/\Psi_2 = \Psi_2/\Psi_3$. But why should we make this assumption? Why not assume instead that equal physical ratios are accompanied by equal psychological *differences*? This would be to assume that: If O estimates that $\Phi_1/\Phi_2 = \Phi_2/\Phi_3$ then $\Psi_1 - \Psi_2 = \Psi_2 - \Psi_3$. On this assumption, Experiment M will establish that the psychophysical law for length is an instance of (2), which is of course Fechner's law. This point is made by Treisman (1964a, pp. 12–16; 1964b, pp. 387–388). The only weakness in his argument is his assumption that Ψ is some neural effect in O, which leaves him open to the criticism made by Stevens (1964, pp. 383–384), who objects that ratio-scaling experiments need not posit intervening neural variables. Treisman's argument can be restated without the fatal assumption, along the lines suggested above.

The point above applies to experiments employing the technique of numerical estimation as well as to those employing the technique of fractionation. Many theorists have distinguished between "direct" and "indirect" psychological measurement. Ekman (1961, pp. 35, 43) says of Stevens' technique: "[I shall call it] *direct* scaling of subjective variables, because the essential steps of the scaling procedure are implied in the experimental situation." Thus when O assigns 100 to the loudness of a standard tone and 62 to that of a comparison, "the two scale values are, by definition, on a ratio scale: the ratio 62/100 should, according to the instructions, be equal to the subjective ratio of the second loudness to the loudness of the standard." Ekman says of Thurstone's methods, and would say the same of Fechner's, that they are "*indirect* methods, since they are based on a set of assumptions intervening between the experimental data and the final scale." Employing this distinction, one might wish to argue that fractionation methods are indirect, since they involve intervening assumptions; whereas numerical-estimation methods involve no such assumptions and are therefore direct. If this were true, then different psychophysical laws could not be derived from numerical-estimation experiments, since there would be no intervening assumptions, no principles of correspondence, to manipulate. But it is not true, for the method of numerical estimation assumes principle of correspondence (Dl): If O assigns to stimuli Φ_1, Φ_2, Φ_3, etc. the numerals m, n, o, etc. respectively, then Ψ_1, Ψ_2, Ψ_3, etc. stand in the ratios $m{:}n{:}o{:}$ etc. This is tacitly admitted by Ekman when he says that scale values in a numerical estimation experiment are "by definition, on a ratio scale." The "definition" in question is just the principle of correspondence stated above. Now other definitions, other principles of correspondence, can as easily be adopted. For instance: If O assigns to stimuli Φ_1, Φ_2, Φ_3, etc. the numerals m, n, o, etc., then the difference between Ψ_1 and Ψ_2 is $m - n$, the difference between Ψ_2 and Ψ_3 is $n - o$, etc. Again, this assumption leads to Fechner's logarithmic law rather than Stevens' power law. The method of numerical estimation is not, in Ekman's sense, direct. (Indeed, no psychophysical method invented is direct in his sense; and probably none could be invented.) Hence, data obtained by employing the method are as subject to arbitrary manipulation as those obtained on any other method.

Where the concept of psychological length is employed, comparisons between different ratio estimates, between different psychophysical functions and between psychophysical and physiologicophysical functions depend entirely on the principles of correspondence employed. But how do we know which principles to use? The choice seems arbitrary, and, consequently, so do the results of the various comparisons. This arbitrariness is not present where we employ methods of comparison which do not employ the concept of psychological magnitude. There-

fore, it seems not merely possible, but also advisable to discard the concept.

THE IMPORTANCE OF THE TWO INTERPRETATIONS

Failure to Distinguish the Interpretations

The present writer is not aware of any place in the literature where the introspectionist and behaviorist interpretations of ratio scales are explicitly and systematically distinguished. Nor do psychophysicists in general appear to observe the distinction in presenting and discussing the results of their experiments. There are two major symptoms of this deficiency: unclarity as to what the *O*s in psychophysical experiments estimate, and unclarity as to what the *E*s in psychophysical experiments measure. It is especially disconcerting, and especially important, to discover such unclarities in the writing of S. S. Stevens, the major architect and methodologist of the so-called "new" psychophysics. If Stevens is unclear, then so is a vast part of contemporary psychophysics. His writings will, therefore, be given most attention.

First symptom. On the behaviorist interpretation *O* in an experiment like M is said to provide quantitative estimates of Φs, that is, physical entities or stimuli, like rods, weights and so on. The *O* is not said to estimate or even be aware of any Ψs or Ψ magnitudes. On the introspectionist interpretation, *O* is thought to make direct quantitative estimates of psychological entities, such as sensations—or at least estimates of psychological magnitudes—in order to make indirect estimates of physical or stimulus magnitudes. Stevens constantly shifts from one of these positions to the other. He mentions all the following as estimated by *O*: (*a*) "the standard stimulus and a set of variable stimuli" (Stevens, 1956, p. 25), (*b*) "the stimuli that arouse two sensory magnitudes" (Stevens, 1956, p. 1), (*c*) "[an attribute of] the stimuli (Stevens & Harris, 1962, p. 489), (*d*) "the apparent magnitude [of the stimulus] *as he perceives it*" (Stevens, 1960c, p. 54), (*e*) "the apparent magnitude [of the stimuli]" (Stevens, 1958b, p. 193), (*f*) "the apparent strength [of the stimuli]" (Stevens, 1960b, p. 239), (*g*) "the apparent strength or intensity of his subjective impressions" (Stevens, 1960b, p. 232), (*h*) "some aspect of his experience" (Stevens, 1956, p. 18), (*i*) "subjective events" or "sensations" (Stevens, 1957, p. 163), (*j*) "sensations" (Stevens, 1954, p. 30; 1956, pp. 24–25), (*k*) "the magnitude of a given sensation" (Stevens, Mack, & Stevens, 1960, p. 64), (*l*) "the relative magnitudes of . . . sensation" (Stevens, 1954, p. 30), (*m*) "attributes of sensation" (Stevens, 1936, p. 406), (*n*) "the apparent intensity of sensations aroused" (Stevens, Mack, & Stevens, 1960, p. 60), (*o*) "the apparent strengths of the sensations produced" (Stevens, 1960b, p. 238).

In (*a*) through (*c*), *O* is said to estimate stimuli or stimulus magnitudes, so a behaviorist interpretation is implied. In (*i*) through (*o*), *O* is said to estimate psychological entities or magnitudes, so an introspectionist interpretation is implied. (*d*) through (*h*) are sufficiently ambiguous to be consistent with either interpretation. Notice that conflicting interpretations are sometimes implied in the same article, sometimes on the same page! Notice also that the vacillation between interpretations continues for a period of three decades to the present.

Stevens might wish to reply that the vacillation is only apparent, the result of using convenient locutions. For in one place (Stevens, 1951, p. 40) he says in describing fractionation experiments:

> . . . a pair of stimuli are given, and the subject estimates the numerical value of their apparent ratio. (More properly stated, he estimates the numerical ratio between the two magnitudes of an attribute of the sensation aroused by the two stimuli, but for the sake of brevity we say simply that he estimates the apparent ratio of the stimuli.)

But this explanation commits Stevens to an introspectionist view of ratio-scaling methods. And it is difficult to believe that he wishes to be so committed.

Galanter (1962, p. 142) provides us with a clear and instructive example of the same vacillation within a single paragraph. He is discussing the category scaling of subjective length.

> The failure of the subject to recognize the re-

peated presentations of the stimuli is not relevant in this category scaling experiment; rather it is his *judgements about the relative magnitudes of the stimuli* that are sought. The experimenter can never decide whether the subject is right or wrong in a scaling experiment; there is therefore no natural way to introduce an outcome structure into this kind of experiment. In scaling experiments we are forced to assume the uncontrolled ability of the subject to *report his sensations*. As we shall see, the reproducibility of the data upon repetition of this experiment lends some support to this assumption. [italics added.]

The occurrence of the italicized phrases shows that Galanter describes the scaling experiment in conflicting ways: first as one in which *O* estimates stimuli, and then as one in which *O* estimates sensations. If *E* cannot decide whether *O*'s estimates are right or wrong, then these estimates cannot be of stimuli, since *E can* decide whether estimates of stimuli are right or wrong. So we seem forced to conclude that *O*'s estimates are of his sensations, since the accuracy of sensation estimates cannot be decided by *E*. But this is to adopt an introspectionist interpretation, which Galanter would presumably disavow. What he should say is that *O* estimates stimuli and that the accuracy of these estimates can be decided by *E*, but that *E* does not attempt to elicit accurate estimates from *O*. The remark that *E* must assume the accuracy of *O*'s estimates is a sure sign that either the writer is an introspectionist or is confused.

Warren & Warren (1963, pp. 804–805) seem to have noticed the vacillation illustrated above.

The New Psychophysics also makes a fundamental distinction between quantitative judgments of subjective (psychological) magnitudes and estimates of physical magnitudes (Stevens, 1958). Thus, in order to construct the *veg* scale of subjective heaviness, *S*s' judgments are considered distinct from estimates of physical weight. However, in the experiment of Harper and Stevens (1948), *S*s were instructed to select that weight "which feels half as heavy as the standard." It is difficult to see how this phrasing differs from instructions to choose an object which seems to be half physical weight. Yet this distinction is essential for Stevens' psychological continua.

These writers are correct in pointing out that Stevens and his co-workers often instruct their observers to estimate physical magnitude, and at other times speak as if judgments of psychological magnitude are being required. But they are wrong in their implication that the new psychophysicists possess an official distinction between judgments of psychological magnitudes and estimates of physical magnitude, and that it is the former which theoretically are required of observers in psychophysical experiments. The distinction seems never to have occurred to most of the psychophysicists in question, and they are far from clear as to what they are or should be requiring from their observers.

Second symptom. Failure to observe the distinction between the introspectionist and behaviorist interpretations also shows up in unclarity as to what experiments like M attempt to measure. On the introspectionist interpretation this sort of measurement consists in assigning numerals to a psychological magnitude privately observed and quantitatively estimated by *O*. On a behaviorist interpretation it consists in quantitatively defining a psychological magnitude in terms of *O*'s quantitative observations of a class of physical stimuli. Again we shall refer to Stevens to illustrate the vacillation between and unclarity regarding these two positions. In a 1936 article (Stevens, 1936, pp. 406–408) we find him saying:

. . . scale numbers [should bear] a reasonable relationship to the experience of the observer. Thus, the scale would be satisfactory if the magnitude of the attribute of sensation to which the number 10 is assigned should appear half as great to the experiencing individual as that to which the number 20 is given, and twice as great as the magnitude to which the number 5 is given. . . . the subjective judgements (responses) of the observer must provide the ultimate test of the validity of the numbers on the scale as representative of degrees of loudness [in general, sensation]. The utilization of the observer's discriminations in this way presupposes, of course, that he is capable of making valid judgements of the numerical ratio of one impression to another.

These passages contain one of the strongest suggestions of an introspectionist view of ratio scales known to the present writer. Stevens says that the scale is satisfactory if "sensations appear . . . to the experiencing individual" to have those ratios which are indicated by the numerals assigned to them by the experimenter; and that the ultimate

test for this result is the "subjective judgements . . . of the observer." To speak in this way is to imply that the entities being measured are private sensations, inner events of which the observer—but not the experimenter—is aware through a faculty of inner perception, through introspection. In addition, the last sentence quoted contains one of the most explicit uses of the assumption that O's sensation estimates are accurate, an assumption which is the second premise in the introspectionist argument, (Ca)-(Cb)-(Cl), discussed earlier.

It is surprising to hear these suggestions from a writer who is known for his attempt to apply the philosophy of operationism to psychology (Stevens, 1939). It is even more remarkable to find these suggestions in the very same pages where Stevens says: "in case of sensations what we want is a scale for the measurement of some aspect of the response of a living organism to a certain class of stimuli" (p. 406), "a subjective scale is a scale of response" (p. 407), and "loudness is a name which we give to a certain class of discriminatory responses" (p. 408). These statements suggest a behaviorist program of defining private sensations, a program laid out in greater detail in the previous year.

> Since sensation cannot refer to any private or inner aspect of consciousness which does not show itself in an overt manner, it must exhibit itself to an experimenter as a differential reaction on the part of an organism. . . . Thus, the *sensation red* is a term used to denote an 'objective' process or event which is public and which is observable by any competent investigator. . . . In the same way that *sensation* denotes a class of reactions which satisfy certain criteria, *attribute* of sensation denotes a sub-class which satisfies more restricted criteria [Stevens, 1935, p. 524].

These passages are puzzling. Stevens says that an E may concern himself with a "private aspect of consciousness" so long as it "exhibits itself . . . as a differential reaction." Isn't this to cling to the private and unmeasureable entity while trying to rid oneself of it? It is one thing to say that sensations which are *defined* as differential reactions can be measured, quite another to say that private sensations which are *exhibited* as differential reactions can be measured. To say the latter suggests that it is the private sensation we really ought to measure—but that since we cannot we can only do second best and measure a public substitute, that is, the differential reaction. Now the pure behaviorist does not cling to private sensations, nor try to measure them in terms of substitutes. He holds that if there are any private entities, they are absolutely of no interest or importance to mathematical science. His position is that certain theoretical constructs may be defined in terms of quantitative observer estimates, which constructs may then be said to be "measured"; and that we may call these constructs "sensations" if we choose. But they are not to be thought of as substitutes or representatives for some sort of private entities. These constructs must stand on their own feet, in virtue of their scientific usefulness, and not in virtue of going proxy for some private entities which the scientist unfortunately cannot observe.

The passages analyzed above are taken from writings early in Stevens' career. It is not unreasonable to expect that in the intervening 30 years he would have clarified his position on the interpretation of ratio scaling. And we do find some evidence of this. In Stevens (1958a, p. 386) he says that since the "non-operational aspects of sensation" are "inaccessible," the "operational stance is indispensable to scientific sense and meaning" in psychology. It follows, he thinks, that "verifiable statements about sensation become statements about responses." In 1959 we can read (Stevens, 1959b, pp. 12, 15) that "immediate experience" with its privacy is not the object of the science of sensation. Sensation, we are told, is, like temperature, "a construct, a conception built upon the objective operations of stimulation and reaction. We study the responses of organisms, not some nonphysical stuff that by definition defies objective test." In 1960 Stevens (1960b, p. 226) cheerfully concedes that we cannot measure the "strength of a sensation" in the "inner, private, subjective" sense; but he insists that there is another sense of "sensation strength" which makes it possible for us to ask "sensible objective questions about the input-output relations of sensory transducers. . ."

These passages clearly suggest a behavior-

ist interpretation of psychophysical measurement. But it is one thing to suggest a position, another to describe it in detail—to lay it out for critical scrutiny and examine its implications. It is, therefore, quite impossible to say what positions Stevens would take on many of the points discussed in this paper. Furthermore, we can still find significant evidence of an introspectionist view of the object of psychophysical measurement in Stevens' later writings. Some of these passages have been provided in earlier sections. One more is worth mentioning (Stevens, 1956, pp. 24–25):

> Sensations do not come with numbers written on them, and when we try to assess the ratio between a pair of them we find ourselves up against a difficult task of appraisal. It is no wonder then that subtle constraints and biases can influence the result. This is another way of saying that the outcome is a function of the method—as it always is in science. What we want, of course, is an unbiased method, one that on the average lets O make an estimate that is neither too high nor too low. Since we do not know in advance what his estimate should be, we can apply no independent criterion of validity.

This passage says that in a ratio-scaling experiment *O* estimates sensations and *E* attempts to assign the correct numerals to these sensations. Stevens implies that both tasks are difficult. This can only be because the sensations have not been previously quantified, which makes it difficult for *O*; and because the sensations are private to *O*, which makes it difficult for *E*. Stevens' lamenting the lack of an "independent criterion of validity" can only be understood by assuming that he regards the scale whose validity is in question as a scale of a private magnitude!

Hirsh (1952, pp. 4–6) provides us with another example of confusion about the object of psychophysical measurement. He begins by noting the difficulty of measuring private entities.

> The end product of our several sensory systems is the sensation—auditory, visual, tactual, olfactory, or gustatory. Each of us knows what a sensation is, because each of us has sensations. . . . It is difficult, however, for each of us to know about another person's sensations, because we cannot get inside his world of experience very easily. You and I may both say 'red' when we see a particular object; but you cannot be sure that my impression of red is exactly the same as yours. . . . we cannot observe the sensations of others, and we can only measure what we can observe.

He then suggests a way out of the difficulty.

> Since we cannot observe the sensation that exists in another individual's world of experience, it would seem indeed that we cannot measure sensation. On the other hand, we can twist the meaning slightly and define the sensation in terms of events that we can measure. When a man says, "I see red," we cannot measure the redness of his visual sensation, nor even be sure that he has one, but we can observe his verbal behavior—"I see red." The phenomena of audition may be studied in the same way. We cannot measure auditory sensations that are private, but we can measure sensations that are defined in terms of behavior or observable responses.

These passages clearly contain a behaviorist-like interpretation of psychophysical measurement. But, in the first place, they are extremely sketchy and vague. How, precisely, are sensations to be defined? In terms of responses to physical stimuli, or in terms of responses to private sensations? If the latter, then the suggestion is still introspectionistic. And how does Hirsh understand "definition"? Does he regard the statements which define sensation as stipulative definitions whose only defense is their usefulness? If not, then his suggestion may still contain introspectionist elements.

In the second place, Hirsh seems, even more clearly than Stevens, to cling to the private entity even while trying to expel it. He laments our inability to "get inside" another individual's world of experience, as if that is what the psychologist really wishes to do. He says that the end products of our sensory systems are private sensations, that we cannot measure these, and that we must settle for measuring something closely related, something defined in terms of behavior or observable responses. The behaviorist does not attempt to measure private entities in terms of public substitutes. His view is that the end products of our sensory systems should be regarded either as physiological processes or as behavioral responses, not as private sensations. The measurement of physiological processes is accomplished, not by procedures like those in Experiment M, but by tapping the organism with instruments, like oscilloscopes. As for behavioral

responses, they may be measured in either of two senses. On the one hand, certain attributes of responses—duration, loudness, etc.—can be measured in the ordinary physical sense by using clocks, meters, etc. On the other hand, certain theoretical constructs can be defined in terms of quantitative behavioral responses, and the entities thus defined can then be said to be measured. But these entities should not be thought of as substitutes for the private entities of which Hirsh speaks.

The Importance of Distinguishing the Interpretations

The previous section was designed to indicate that the distinction between the introspectionist and behaviorist interpretations of ratio scales has more than academic interest. In spite of its importance, there is almost universal failure even to suggest the distinction, much less explicitly to draw it. And there is a universal tendency, found even among those who mention the distinction, to run the two interpretations together. The explanation for this confusion may be as follows. The most natural and attractive way of viewing ratio scaling in particular and psychophysical measurement in general is to see it as the attempt to quantify privately observable, empirically real magnitudes. But the consequences of this view are that psychological magnitudes cannot be measured nor psychophysical laws verified in an acceptable scientific manner. The philosophically sensitive psychophysicist recognizes these consequences, and seeks to avoid them by adopting a behaviorist interpretation of psychophysical measurement. He concedes that although the private psychological magnitudes are incapable of scientific treatment, psychological magnitudes as defined in terms of observer responses can be measured and laws regarding them verified.

However, it is difficult to discard the introspectionist interpretation completely, and for more than one reason. In the first place, this interpretation is the more natural and attractive of the two, both because it has a longer history and also because it is, for some deeper reason, intellectually the most satisfying. Secondly, the psychophysical experimenter tends to be impatient with philosophical and methodological considerations, and to be unwilling to lay the behaviorist definitions out in detail and to examine their implications with care. But more fundamental than either of these reasons is the logical similarity between the theoretical fictions of the behaviorist interpretation and the empirically real, observable psychological magnitudes of the introspectionist.

Although psychological length as defined in (i) and (ii) is merely a shadow of its introspectively observable counterpart, it is still a magnitude: one can still speak of Ψs as being greater than, equal to, and less than one another. And although the principle of correspondence is, on the behaviorist view, merely a stipulative definition, still it receives the same formulation as the introspectionist principle. These similarities produce a strong tendency to slip from the behaviorist interpretation back into the introspectionist, or, what is virtually the same error, to treat psychological length as defined in (i) and (ii) as a substitute, or proxy, for the privately observable, scientifically unmanageable psychological magnitude posited by the introspectionist. This tendency manifests itself in a number of ways, one of the most important of which is to think of psychological length as a *single* magnitude which observers may estimate in a number of different ways. Thus *O*'s one-half and one-fourth estimates in Experiment M are thought of as different fractional estimates of entities which lie along a single continuum. And jnd, interval, and ratio scales are thought of as different scales of a single magnitude, so that if neither scale agrees with any of the others then only one of them can be "valid."

Failure to distinguish the introspectionist from the behaviorist interpretation produces vacillation between the two, and serves to conceal the difficulties in both. Psychological length is on the introspectionist interpretation an "empirically real" but scientifically unmanageable magnitude; on the behaviorist a scientifically manageable but "fictional" magnitude. Running the two interpretations together creates the delusion that psychological length is both "empirically

real" and scientifically manageable. On the introspectionist interpretation Psychophysical Law (5) is a description of *O*'s perceptual mechanism; but it is unverifiable since only *O* can observe Ψs. On the behaviorist interpretation (5) is verifiable in the standard scientific manner, but it is not an explanation of how *O* perceives physical length. By amalgamating the two interpretations, (5) seems to emerge both as publicly verifiable and as an explanation of *O*'s perceptual behavior. But the amalgam is, of course, unstable, since it incorporates contradictory elements.

Only when we carefully distinguish the two interpretations, laying each out in detail as we have done in previous sections, does it become possible to assess ratio-scaling procedures. It seems clear that the mak scale is illegitimate if interpreted introspectionistically. The *O* observes no private psychological entities; and even if he did, the mak scale would provide us, not with measurements, but only with estimates of these entities. Furthermore, Psychophysical Law (5) is unverifiable on this interpretation: both because it rests on principle of correspondence (Cl), which in turn rests on unverifiable assumptions (Ca) and (Cb); and also because *O*'s (putative) quantitative estimates of psychological length are unconfirmable. (We must say "unverifi*ed* assumptions" and "unconfirm*ed* estimates" where *O* is thought to estimate a physiological magnitude.)

If we adopt a behaviorist interpretation, these positive objections no longer apply. Rather the objection becomes the negative one that the scaling of psychological length is unnecessary, because the concept of psychological length is unnecessary. In a previous section it was tentatively argued that any legitimate purpose in constructing a ratio scale of psychological length can be equally well achieved without employing the concept of psychological magnitude. If this is so, then the concept is dispensable and *may* be abandoned. And there are reasons for thinking that it *should* be abandoned. One of these is simply Occam's maxim, which says: Do not multiply entities beyond necessity. Put in more contemporary fashion: Do not clutter up the theoretical system with unnecessary constructs. If this reason seems insufficient, we may point out the confusion caused by the constant temptation, described earlier, for the behaviorist to slip back into an unacceptable, introspectionist way of thinking about psychological magnitudes. Other reasons emerge by considering alternatives to those systems which include the concept of psychological magnitude.

Beyond the Two Interpretations

It may seem that to abandon the concept of perceptual magnitude is to abandon perceptual psychophysics. This is not so. It is rather to adopt a view of the nature of perceptual psychophysics which differs both from that of the "old" psychophysics of Fechner and the "new" psychophysics of Stevens. Whatever their differences, both theorists view perceptual psychophysics as the attempt to measure perceptual magnitudes in order to discover mathematical laws relating these to stimulus magnitudes. So it is clearly unorthodox to recommend doing away with the concept of perceptual magnitude. But it is one thing to embrace an unorthodox view of the nature of perceptual psychophysics, and another to abandon the science. The most promising unorthodox view would replace the concept of a perceptual magnitude with that of a *perceptual*, or discriminatory, *ability*. Perceptual psychophysics then becomes the experimental discipline which describes, measures, predicts, and perhaps even explains the perceptual abilities of organisms. Let us examine the consequences of applying this view of the science to Experiment M.

M is really an attempt to determine *O*'s ability to make fractional estimates of rod length by sight. The *O*'s ability to do this may be perfect, or it may be less than perfect in varying degrees. These degrees of ability can be represented in psychophysical functions of the form, $\Phi_c = b\,\Phi_s$. If *O* were able to quarter lengths perfectly the value of b would be .25. If he were able to halve lengths perfectly the value would be .50. To the extent that b falls below this value, *O* underestimates the comparison rod (overestimates the standard). To the extent that b exceeds this value, *O* overestimates the comparison rod (underestimates the

standard). Thus the value of b provides a measure of O's ability to make fractional estimates of length. This measure can be obtained merely by constructing the Φ-Φ functions of Figure 1. The Ψ-Φ functions of Figure 2 are not only unnecessary; it is difficult to see immediately what connection they have with perceptual ability.

The discovery of Φ-Φ functions makes it possible for us to compare O's abilities to make fractional estimates of one-fourth with his ability to make fractional estimates of one-half, one-third, two-thirds, etc. O overestimates the comparison rod in the first part of Experiment M by .34–.25/.25, or 36%. He overestimates the comparison rod in the second part by .58–.50/.50, or 16%. This makes it possible for us to say not only that O's ability to fourth lengths is not as great as his ability to halve them, but also how much greater the one is than the other. What this difference in abilities shows is not our concern here. But one might wish to consider the hypothesis that other things than length are among O's cues, or that he does not comprehend fractions perfectly, or even that different physiological mechanisms are involved in the two fractionations. As we saw in an earlier section, the Φ-Φ functions also make it possible for us to compare O's ability to make ratio estimates with his ability to make interval estimates of length. Again, as we saw previously, by comparing psychophysical Φ-Φ functions with their associated physiological θ-θ functions, we can make inferences regarding the physiological mechanisms underlying O's ability to make ratio estimates.

The point is not that perceptual capacities can be measured and studied only through the discovery of psychophysical Φ-Φ functions. Functions which relate number of psychological intervals to stimulus magnitude (like the jnd function in Figure 2) are also useful. And still other sorts of function, as well as a variety of statistical methods, can be employed by E. We merely wish to suggest that in perceptual psychophysics every legitimate question can be raised and answered, and every legitimate comparison and inference can be made, by employing the concept of a perceptual ability. If this is so, then perceptual psychophysics can be just as rich, interesting, and productive when construed as the science of perceptual abilities as it is when construed, in the manner of Fechner and Stevens, as the science of perceptual magnitudes. Furthermore, our unorthodox view has the positive advantage of laying to rest those nagging philosophical doubts about the possibility of psychological measurement which have attended psychophysics since Fechner founded the science.

There is no philosophical problem of psychological measurement for the science of perceptual abilities. On the orthodox view, psychological measurement is regarded as the measurement of psychological magnitudes. But there is a question as to whether such measurement is possible. If psychological magnitudes are private, as the introspectionist believes, then they are, it would appear, incapable of measurement. If we define psychological magnitudes in terms of observer responses, as the behaviorist does, then they appear capable of measurement. But this solution seems to allow the prize to slip through our fingers. What we really wished to measure was an empirically real, although private, dimension of mind; but all we managed to measure was an anemic substitute, a theoretical construct or fiction. We really wanted to discover the relation between a psychological magnitude correlated with but defined independently of a physical magnitude; but instead we were forced to define the psychological magnitude in terms of (observer responses to) physical magnitude.

These difficulties are completely circumvented by construing psychological measurement as the measurement of perceptual abilities. For when we abolish psychological magnitudes, such as psychological length, psychological weight, etc., then there can be no question of how we measure them. Viewed as an attempt to determine ability to fractionate length, Experiment M consists of (*a*) measuring the lengths of the rods to be presented, (*b*) eliciting quantitative estimates of rod length from O, and (*c*) computing the Φ-Φ functions in Figure 1. Nowhere in these three stages do we find any attempt to measure psychological magnitudes. Speaking loosely, we may say that (*a*), (*b*), and (*c*) constitute a complex proce-

dure for measuring *O*'s ability to fractionate physical length. But strictly speaking, the only measurement which occurs in M is the measurement of physical length in stage (*a*). And since there is no problem in measuring physical length, there is no problem of measurement when M is regarded as an attempt to determine one of *O*'s perceptual abilities.

In the "new" psychophysics the old philosophical doubts about the possibility of psychological measurement reappear with a somewhat different emphasis, that is, as doubts about the possibility of determining which of the various types of psychological scale is "valid." In the attempt to scale perceptual magnitudes such as psychological length, a problem arises over the validity of competing scales. Shall we employ a jnd, an interval, or a ratio scale? On the introspectionist interpretation this problem is unavoidable and insoluble. Each scale is constructed from what are taken to be *O*'s estimates of private psychological entities. Since there is no way of confirming such estimates, there is no way to determine whether the scales built from them accurately represent the magnitude of the private entities. On the behaviorist interpretation, there still seems to be a problem. Psychological magnitudes are defined in terms of *O*'s publicly observable responses to physical stimuli, so that the privacy problem is solved. But do *O*'s jnd responses, his interval responses, and his ratio responses to length define three different psychological magnitudes, three different psychological lengths? It seems odd to say that they do, since they are all responses to the same physical magnitude. On the other hand, if a single psychological magnitude is involved, then scales constructed from the different responses compete with one another and we must choose between them. And how shall we do that?

Where we take ourselves to be scaling perceptual abilities, these difficulties do not arise. There is no philosophical problem concerning the validity of psychological scales for the science of perceptual abilities. It is obvious that a scale constructed from the partition estimates of an observer can at best provide a measure only of his ability to make partition estimates: it cannot provide a measure of his ability to make ratio estimates. To deny this would be like asserting that a scale of a person's muscular ability to lift rods can provide a measure of his ability to discriminate rods, or that a scale of his reading ability can provide a measure of his mathematical ability. Of course it may be necessary to choose between different measures of perceptual ability. For example, after discovering the Φ-Φ functions of Figure 1, we must decide whether *O*'s ability to fractionate lengths is best represented by the *difference* between the obtained value of b and the value for a perfect estimate (.25 for quartering, .50 for halving) or by the *ratio* between these values. But this decision can be made on the basis of scientific convenience and usefulness. It involves no philosophical problems.

REFERENCES

BERGMANN, G., & SPENCE, K. W., The logic of psychophysical measurement. *Psychological Review*, 1944, **51**, 1–24. Reprinted in H. FEIGL & M. BRODBECK (Eds), *Reading in the philosophy of science*. New York: Appleton-Century-Crofts, 1953. Pp. 103–119.

BORING, E. G. The stimulus error. *American Journal of Psychology*, 1921, **32**, 449–471.

BORING, E. G. Did Fechner measure sensation? *Psychological Review*, 1928, **35**, 443–445.

BORING, E. G. The psychologist's circle. *Psychological Review*, 1931, **38**, 177–182.

EKMAN, G. Two generalized ratio scaling methods. *Journal of Psychology*, 1958, **45**, 287–295.

EKMAN, G. Some aspects of psychophysical research. In W. A. ROSENBLITH (Ed.), *Sensory communication*. New York: Wiley, 1961. Pp. 35–47.

EKMAN, G. T., & SJÖBERG, L. Scaling. *Annual Review of Psychology*, 1965, **16**, 451–474.

FECHNER, G. *Elements of psychophysics*. Vol. I. Translated by Helmut E. Adler. New York: Holt, Rinehart, & Winston, 1966.

FERGUSON, A., et al. Quantitative estimates of sensory events. Final report of a special committee. *British Association for the Advancement of Science*, 1939–1940, 331– 49.

GALANTER, E. Contemporary psychophysics. In

ROGER BROWN (Ed.), *New directions in psychology*. New York: Holt, Rinehart & Winston, 1962. Pp. 87–156.

GARNER, W. R. A technique and a scale for loudness measurement. *Journal of the Acoustical Society of America*, 1954, 26, 73–88.

GIBSON, J. J. *The perception of the visual world*. Boston: Houghton Mifflin, 1950.

GOUDE, G. *On fundamental measurement in psychology*. Stockholm: Almquist & Wiksell, 1962.

GUILFORD, J. P. *Psychometric methods*. (2nd ed.) New York: McGraw-Hill, 1954.

GUILFORD, J. P., & DINGMAN, H. F. A validation study of ratio-judgment methods. *American Journal of Psychology*, 1954, 67, 395–410.

HARPER, R. S., & STEVENS, S. S. A psychological scale of weight and a formula for its derivation. *American Journal of Psychology*, 1948, 61, 343–351.

HIRSH, I. J. *The measurement of hearing*. New York: McGraw-Hill, 1952.

JOHNSON, H. M. Did Fechner measure 'Introspectional' sensation? *Psychological Review*, 1929, 36, 257–284.

PRICE, H. Review of Gibson, J. J. *The perception of the physical world*. *Mind*, 1953, 62, 406–410.

REESE, E. P., REESE, T. W., VOLKMANN, J., & CORBIN, H. H. (Eds) *Psychophysical Research Summary Report*. NAVEXOS P-1104. SDC 131-1-5, 1953.

REESE, T. W. The application of the theory of physical measurement to the measurement of psychological magnitudes. *Psychological Monographs*, 1943, 55, (3, Whole No. 251).

SAVAGE, C. W. The measurement of loudness and pitch. Unpublished doctoral dissertation, Cornell University, 1963.

STEVENS, J. C., MACK, J. D., & STEVENS, S. S. Growth of sensation on seven continua as measured by force of handgrip. *Journal of Experimental Psychology*, 1960, 59, 60–67.

STEVENS, S. S. The operational definition of psychological concepts. *Psychological Review*, 1935, 42, 517–527.

STEVENS, S. S. A scale for the measurement of a psychological magnitude: Loudness. *Psychological Review*, 1936, 43, 405–417.

STEVENS, S. S. Psychology and the science of science. *Psychological Bulletin*, 1939, 36, 221–263. Reprinted in M. MARX (Ed.) *Psychological Theory*. New York: Macmillan, 1961. Pp. 21–54.

STEVENS, S. S. Mathematics, measurement, and psychophysics. In S. S. STEVENS (Ed.), *Handbook of experimental psychology*. New York: Wiley, 1951. Pp. 1–49.

STEVENS, S. S. Biological transducers. *Convention record of the Institute of Radio Engineers*, 1954, Part 9, 27–33.

STEVENS, S. S. The direct estimation of sensory magnitude—loudness. *American Journal of Psychology*, 1956, 69, 1–25.

STEVENS, S. S. On the psychophysical law. *Psychological Review*, 1957, 64, 153–181.

STEVENS, S. S. Measurement and man. *Science*, 1958, 127, 383–389. (a)

STEVENS, S. S. Problems and methods of psychophysics. *Psychological Bulletin*, 1958, 55, 177–196. (b)

STEVENS, S. S. Measurement, psychophysics, and utility. In C. W. CHURCHMAN & P. RATOOSH (Eds), *Measurement: Definitions and Theories*. New York: Wiley, 1959, Pp. 18–63. (a)

STEVENS, S. S. The quantification of sensation. *Daedalus*, 1959, 88, 606–621. (b)

STEVENS, S. S. On the validity of the loudness scale. *Journal of the Acoustical Society of America*, 1959, 31, 995–1003. (c)

STEVENS, S. S. On the new psychophysics. *Scandinavian Journal of Psychology*, 1960, 1, 27–35. (a)

STEVENS, S. S. Psychophysics of sensory function. *American Scientist*, 1960, 48, 226–253. Reprinted in W. A. ROSENBLITH, (Ed.), *Sensory Communication*. New York: M. I. T. Press and Wiley, 1961. Pp. 1–33. (b)

STEVENS, S. S. Ratio scales, partition scales, and confusion scales. In H. GULLIKSEN & S. MESSICK (Eds.), *Psychological scaling: Theory and applications*. New York: Wiley, 1960. Pp. 49–66. (c)

STEVENS, S. S. To honor Fechner and repeal his law. *Science*, 1961, 133, 80–86.

STEVENS, S. S. The surprising simplicity of sensory metrics. *American Psychologist*, 1962, 17, 29–39.

STEVENS, S. S. Concerning the psychophysical power law. *Quarterly Journal of Experimental Psychology*, 1964, 16, 383–385.

STEVENS, S. S., & DAVIS, H. Psychophysiological acoustics: Pitch and loudness. *Journal of the Acoustical Society of America*, 1936, 8, 1–13.

STEVENS, S. S., & GALANTER, E. H. Ratio scales and category scales for a dozen perceptual continua. *Journal of Experimental Psychology*, 1957, 54, 377–411.

STEVENS, S. S., & HARRIS, J. The scaling of subjective roughness and smoothness. *Journal of Experimental Psychology*, 1962, 64, 489–494.

STEVENS, S S., & STONE, G. Finger span: Ratio scale, category scale, and jnd scale. *Journal of Experimental Psychology*, 1959, 57, 91–95.

STEVENS, S. S., & VOLKMANN, J. The relation of pitch to frequency: A revised scale. *American Journal of Psychology*, 1940, 53, 329–353.

TITCHENER, E. G. *Experimental psychology*. Vol. 2, Part I. *Student's manual*. New York: Macmillan, 1905.

TREISMAN, M. Sensory scaling and the psychophysical law. *Quarterly Journal of Experimental Psychology*, 1964, 16, 11–22. (a)

TREISMAN, M. What do sensory scales measure? *Quarterly Journal of Experimental Psychology*, 1964, 16, 388–391. (b)

WARREN, R. M., SERSEN, E. A., & PORES, E. B. A basis for loudness-judgments. *American Journal of Psychology*, 1958, 71, 700–709.

WARREN, R. M., & WARREN, R. A Critique of Stevens' 'New Psychophysics.' *Perceptual and Motor Skills*, 1963, 16, 797–810.

(Received April 28, 1966)

OPERATIONS OR WORDS?

S. S. STEVENS

Laboratory of Psychophysics, Harvard University

A comment on Psychological Monograph 627, Part I, by C. Wade Savage, Introspectionist and Behaviorist Interpretations of Ratio Scales of Perceptual Magnitudes.

WHENEVER his choice of words causes misunderstanding, the operational scientist ought perhaps to invert the well-known paternal dictum and advise his readers to follow what he does, not what he says. A chief leverage of the operational stance resides in its ability to sustain the concepts of science, regardless of the vagaries of changing styles in verbal behavior. Under the drive to keep abreast of the evolving, turbulent jargon of science, we alter our habits of speech with less and less concern for tradition, but the useful meanings of words continue to rest with the operations to which the words refer. If that point of view, so basic to science and so natural to most scientists, had been adopted by C. Wade Savage (1966), perhaps his scholarly review of the ratio scaling of perceptual magnitudes would have assumed a more mellow cast. A good operationist, I think, would have shown little shock and alarm concerning my many verbal vacillations as I have sought over the years to convey to different audiences a feeling for the goal of psychophysics. Yes, if Savage would attend to what I do and not to what I say, there is a good possibility that he might come to share an enthusiasm for the psychophysical power law and the several scientific questions it has illuminated. It is what we do in the experiments that matters most, not what words we use to describe our results. For words are made to serve us, not to rule over us.

Savage has produced an erudite and forensic endeavor that deserves to be scrutinized with care, as he himself has scrutinized his many references. His alarm at my verbal transgressions is only incidental. The broader purpose is to examine the implications of two opposing points of view, called introspectionist and behaviorist. They are idealized points of view, not necessarily representative of any psychophysicist, living or dead. The two views are verbal creations, each posing as an active agent that is able to make interpretive statements about psychophysical experiments. Thus the introspectionist says that psychological magnitudes are observable but private; whereas the behaviorist says that they are public but nonobservable. (I myself would say that they are public and observable, which would serve to justify one of Savage's complaints, that I refuse to fit neatly into his categories.)

The fabricating of a straw-filled protagonist—or even a pair of them—can provide good entertainment as well as an occasion for the clarification of meanings at the syntactical level. Since there exist few natural constraints on the stuffing of verbal straw, the introspectionist and the behaviorist can be pictured as extreme types, exaggerated and caricatured, true to no actual man. They may then become personae through which the author speaks, and since they can be cast in antithetical roles, their verbal jousting can have the form if not the substance of empirical debate. There need be no harm in such logomachy, unless perhaps the word game happens to deceive its own creator. That, in fact, is what may have occurred.

Having sketched the models of the two "interpretations," introspectionist and behaviorist, Savage turns his attention to the phrasings of various psychophysicists and finds them wanting. He who writes about a psychophysical experiment may be hanged, we are led to believe, merely for the words he uses. If he says the listener judged the apparent magnitude of sensation, he is guilty of the introspectionist interpretation. If he says the listener compared a variable

stimulus to a standard stimulus, he commits himself to the behaviorist view. Since I myself, in striving to effect a reasonable degree of communication with other people, have used many different phrasings, I am said to writhe in conflict and to vacillate in indecision. I beg to plead guilty on both counts, but not for the reasons given; my conflicts and vacillations usually concern matters of greater substance than mere words. But that is just the point, what consequence hinges on the words we use? A rose by any other name. . . .

The prime reason for using care in the choice of words is to smooth the flow of communication. We want if possible to use the word that correctly taps the reader's semantic set and speeds him to a grasp of the operations referred to. Replying to my Aunt Emma's query about what I do at Harvard besides teach, I tell her that I measure people's sensations—for instance, what they hear and how loud it is. She once answered, "Yes, I had my hearing measured recently." I let it go at that. A description in terms of the measuring of sensations, particularly the loudness of noises, is also a phrasing that may speed communication with an acoustical engineer. With some of my philosopher friends, however, I prefer to say that I try to determine the relative responses of sensory systems by means of cross-modality matching. If I use the words sensation or subjective, some philosophers leap to the verdict that my task of measurement is impossible on the face of it. As these examples suggest, it is the reader's semantic set that determines how a phrase will sit, and the writer is usually powerless to mold his reader's vocabulary into some preconceived ideal. So the writer casts about for the apt description, and on different occasions he is likely to say it in different ways.

I am impressed by the care with which Savage has tabulated examples of my behavioristic warp and my introspectionistic woof. My variations in phrasing are found to be symptoms of "unclarity." My thesaurus seems to have failed me. I am said to vacillate back and forth. Actually, it seems to me that the tabulation of my phrasings shows evidence of a drift, a kind of upward trend toward freer expression. It had not been apparent to me before, but the passing of three decades has indeed made a difference. When behaviorism was new and still self-conscious about its words, there were taboos and prohibitions, as though we could lay ghosts merely by rubbing out their names. But the development of operationism, aided by the semantic analyses contributed by some of the logical positivists, has led to the view that verbal taboo, like word magic of all sorts, is unworthy of the scientist. It is not what we say, it is what we mean that counts, and to get at those meanings we must go behind the words to the operations indicated. The scientist in his contest with nature depends not on words but on operations—methods, procedures, inventions, devices, manipulations, the whole paraphernalia of science. Now that I have come to heed the counsel of the operational view, whenever I measure subjective sensation I am apt to say so in just those words. What do I mean? Let us consider briefly a pardigm of the operations that are intended, a paradigm discussed at greater length elsewhere (Stevens, 1966a).

The Matching Operation

In the study and measurement of perceptual magnitudes we utilize acts of judgment as one of the basic operations. But what is judgment other than a process of comparing, equating, or matching? Consider an example. In measuring luminance, the photometrist changes the brightness of the comparison field until he judges it to be equal to the brightness of the target field. He behaves as a comparator. If you alter the target field, he will adjust the comparison field accordingly. His judgment is clearly a matching operation. If we extend this matching paradigm to other kinds of judgment, we find that the core of all judging operations is a matching—a coupling or conjoining of an element from one domain with an element from another domain. It is hard to envisage an act of judgment that cannot be so construed.

Next let us consider the particular type

of judgment that is basic to the measurement of sensation. A scale of loudness can be constructed by asking observers to produce a match between loudness and perceptual values on one or more other continua. Many such cross-modality matches have been made, and the results have been mapped into a self-consistent family of functions.

Examples of such functions are shown in Figure 1, where the results obtained by matching loudness to 10 other continua are displayed in log-log coordinates (Stevens, 1966b). Perhaps the most provocative feature of the matching functions is their transitivity: any two matching functions can be used to predict a third. Thus the slopes of two of the lines, for example, those relating loudness to handgrip and loudness to brightness, allow us to predict the slope of the matching function when observers make a direct match between handgrip and brightness.

Any one of the functions in Figure 1 could be used to define the loudness scale, the so-called sone scale, for the choice of a reference function, like the choice of a unit, is basically arbitrary. Contrary to what is sometimes assumed, the matching of loudness to number enjoys no privileged status in the family of matching functions. It is merely another cross-modality function. If you take that function away, where are we left? Essentially right where we were. We can then define the sone scale in terms of matches between loudness and some other continuum, apparent length, say. All the engineering applications for which the sone scale has proved useful would still be open to us.

I have discussed the creation of so-called subjective scales in terms of the matching operation because it seems to present a better paradigm than the example offered by Savage, which was the scaling of apparent length by means of the now obsolescent procedure of fractionation. Fractionation also involves a matching process, as I have noted elsewhere (Stevens, 1966a), but, since the matching involves prescribed ratios, fractionation exhibits complexities not shared by direct cross-modality matching. As a consequence, discussions of fractionation have often swerved toward tangential issues, issues that obscure the prime concern. The focal issue centers on the use of the observer as a comparator. We measure a process in the observer, such as perceived magnitude, by systematic study of his behavior as a matching comparator—his behavior as a balancer, equater, or conjoiner.

Perceived magnitude, sensory intensity, or whatever you care to call it, becomes, then, a construct. But that is not surprising, because all useful empirical concepts are constructs. Concepts vary, of course, in the specificity of their operational basis, and the direction of progress for many scientific constructs is toward increasingly precise and determinate definition. The concept of perceptual magnitude rested initially on the operations of verbal response (matching with adjectives), exemplified by such phrases as strong taste, faint smell, blinding light, soft sound, and so forth. With the invention of devices to control stimuli, it became possible to turn the level of a sound up and down and to note

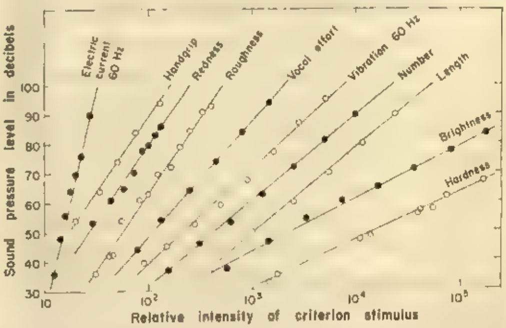

FIG. 1. Examples of equal-sensation functions obtained by matches between loudness and perceptual values on various other continua. The straight lines through the data define power functions in the log-log coordinates. The slope of the line gives the exponent of the power function. In most of the 10 experiments, groups of 10 or more observers adjusted the level of a sound to match criterion stimuli presented in irregular order in another sense modality. For handgrip, vocal effort, and length, fixed sound levels were set, and the observer varied the other stimuli to match the loudness. For the function labeled number, the listeners match loudness to a series of numbers spoken by the experimenter.

the apparent change in loudness. It was promptly observed that a logarithmic increase in sound level did not seem to produce a linear increase in loudness, as many had thought it should. Thereupon some of the pioneering work on loudness scaling was undertaken to determine the relation between loudness and sound pressure level. The functions in Figure 1 represent an example of a later stage in the continuing effort to specify with ever increasing certainty the function that relates subjective loudness to stimulus intensity.

Perceptual Magnitude

Savage's dismay at the failure of some of us to use verbal forms that would consign us consistently to one or the other of his categories of "interpretation" seems to originate from a belief that words mean only what he happens to think they mean, not what some of the rest of us think they mean. That is a very human tendency, of course, but one that may lead to cross-purposes when a man's choice of a word is mistaken for his "way of thinking." Thus when I speak of psychological or subjective magnitude, I do not refer to the scientifically unmanageable private entities so dear to the introspectionist, as posited by Savage. I refer, rather, to a configuration of defining operations, a configuration that is growing more and more definitive as functions like those in Figure 1 become better determined.

It is interesting to note that the physicist's definition of the physical intensity of a sound has also changed rapidly in recent years. Half a century ago, it was not possible to measure such parameters of an acoustic wave as the sound pressure. The operations had not been invented. With the development of the microphone and the vacuum-tube amplifier it became for the first time possible to record the minute pressure changes produced by ordinary sound waves. As knowledge in this area has increased, the operational definitions have become more precise, but none of the definitions are as yet petrified in rigidity, nor are they beyond further improvement.

Savage proposes that the concept of perceptual magnitude is dispensable. I feel sure that the words themselves are dispensable, but is that what he means? Or does he mean that all matching functions, such as those in Figure 1, must be filed in the wastebasket? If the concept is to be abandoned, what do we do with the configurations of matching operations on which the concept rests?

Savage complains that my writing lacks clarity. I do indeed fall short in my aspiration to compose impeccable prose, but I also have another frailty: I sometimes fail to decipher precisely what another writer has in mind. Savage says that we should replace the concept of perceptual magnitude with another concept, namely, perceptual ability. My failure here is to understand whether the proposal is for a verbal substitution or whether it augurs a substantive alteration in my laboratory experiments. As a corollary, the question presents itself whether the matching functions in Figure 1 are to be construed as measures of perceptual ability. If so, then we are engaged in a verbal substitution, and the question arises whether the term ability is a wise choice to convey our meaning. One is reminded of the debate that has swirled about the concept of mental ability. Is psychophysics about to plunge into a similarly sticky morass? Does ability mean a potential for performance, or merely what a person happened to do on a given occasion? Maybe what he did was an accident and unrepeatable. What then? Contemplation of such questions may well suggest that perceptual magnitude is quite as amenable to operational definition as is perceptual ability.

Bits and Pieces

There are several minor queries that suggest themselves in Savage's essay. I shall try to note only a few of them.

Does the new psychophysics try to distinguish between estimating subjective magnitude and estimating physical magnitude? Yes, it does. From an experienced photographer you can get two estimates of the light level, depending on whether you ask him to judge the physical or the apparent level. And for the apparent

or subjective value he will give widely different estimates if you vary the state of adaptation of his eyes. Similarly, a sound engineer can estimate the decibel level of a noisy factory, or, under a different *Aufgabe,* he can judge apparent loudness. In an experiment on the judgment of visual areas, Martha Teghtsoonian (1965) found that magnitude estimations of apparent size gave a power function with an exponent of 0.76. When asked to judge the physical area, the observers changed their basis of judgment, as evidenced by the exponent's increasing to 1.03.

In view of the foregoing facts, I do not understand why Savage seems to say that "the new psychophysicists [do not] possess an official distinction between judgments of psychological magnitude and estimates of physical magnitude." There is nothing particularly "official" about it, but some experimenters are keenly aware of the distinction and try to be careful to state it—whenever it matters. When it does not matter, as it often does not, then an observer may be instructed merely to match stimulus x to stimulus y, or some such task. If Savage would try running a number of these psychophysical experiments, he would probably discover the virtue of framing his instructions to the observers with an eye to ease of comprehension. Thus with an attribute like length, it matters little whether a person is told to judge the subjective length, the length, or simply the lines. Nor is it necessary to tell him to ignore the considerable manifold of other attributes, such as apparent thickness, color, tilt, duration of presentation, etc., any one of which could be judged if the experimenter happened so to instruct the observer. On the other hand, a careful wording of the instructions is probably needed when the purpose is to elicit matching judgments for auditory density or auditory volume (Stevens, Guirao, & Slawson, 1965).

My ability to communicate with Savage is perhaps nowhere more disappointing than on the topic of measurement. Somehow I sense in his discussion of measurement the atmosphere of a debating contest in which the opposition is considered wrong on all issues, regardless. The measure of satisfaction that I had sometimes felt at having introduced the clarifying concept of invariance into the theory of scales is somewhat shattered by the manner in which Savage gives that concept the back of his hand. I still hope it deserves better, and that the theory of scale types will continue to be characterized in terms of the group of transformations that leaves a scale form invariant.

One final word about words. The use of stimulus-response language does not make a man a behaviorist, nor does the use of terms like subjective, or even private experience, make a man an introspectionist. It is hardly that simple. You cannot tell them solely by their spots, at least not by the spots they make on paper. A scientist who tries to adhere to the operational view will insist on looking behind the words to discover the occasions and circumstances of their use. The term subjective may then turn out to mean merely that a living subject participated in the experiment—animate rather than inanimate. In a similar vein, private experience may mean only that the reaction in question occurred in one particular person. In a similar sense, each of the ammeters on the bench before me enjoys its own private experience when its working element (a coil suspended in a magnetic field) performs the comparator task of indicating the point at which a torque produced by magnetic forces balances a torque produced by elastic forces. Different ammeters behave in different ways, and I can study their matching behavior much as I study the matching behavior of people, namely, by mapping the input-output matching functions. Each person's matching judgment in a psychophysical experiment is a private affair, but, in a comparable sense, the comparator process in a given ammeter is the private affair of that instrument.

In conclusion, I should like to press a precept that seems acutely relevant to the study of perception. When we study the input-output characteristics of ammeters, we do not feel called upon to imagine how it feels to be an ammeter, nor do we try to relate our own experiences to those of

ammeters. In the scientific study of man, especially in the study of the operating characteristics of his sensory systems, many pseudo problems can be bypassed if we take the same objective attitude toward the human participant in an experiment as we take toward an ammeter. We may then carry out psychophysical experiments with human subjects much as we perform them with animal subjects, claiming no privileged view of things merely because we, as experimenters, happen to be human. (If some animal should decide to experiment on human beings, it ought, no doubt, to try to follow a similar precept.) In our actual, concrete experiments we do, of course, proceed as though the human subject is an object of study. It is usually when we set about to describe what we have done that troubles arise, for the layers of meanings that attach to the words we use may then mislead some of our readers, even as I seem to have misled Savage.

REFERENCES

SAVAGE, C. W. Introspectionist and behaviorist interpretations of ratio scales of perceptual magnitude. *Psychological Monographs*, 1966, **80** (15, Whole No. 627, Part 1).

STEVENS, S. S. On the operation known as judgment. *American Scientist*, 1966, in press. (a)

STEVENS, S. S. Matching functions between loudness and ten other continua. *Perception and Psychopyhysics*, 1966, **1**, 5–8. (b)

STEVENS, S. S., GUIRAO, MIGUELINA, & SLAWSON, A. W. Loudness, a product of volume times density. *Journal of Experimental Psychology*, 1965, **69**, 503–510.

TEGHTSOONIAN, M. The judgment of size. *American Journal of Psychology*, 1965, **78**, 392–402.

(Received June 21, 1966)

Vol. 80, No. 20 Whole No. 628, 1966

Psychological Monographs: General and Applied

PARTIAL REINFORCEMENT EFFECTS WITHIN SUBJECT AND BETWEEN SUBJECTS[1]

ABRAM AMSEL, MICHAEL E. RASHOTTE, AND JOHN R. MACKINNON[2]

University of Toronto

Comparisons between the designs of within- and between-S partial reinforcement experiments were made as a prelude to 4 experiments in which rats were given acquisition and extinction training in black and white straight runways. For within-S training 1 runway was associated with partial reinforcement (PRF) and the other with continuous reinforcement (CRF). Between-S groups were also trained in the two runways, the CRF group receiving continuous reward and the PRF group partial reward in both. Only within-S groups were trained in the first two experiments. In Experiment 1 rewards, and in Experiment 2 trials, in the PRF and CRF runways were equated. The acquisition data of both experiments were similar to previous data showing faster terminal PRF speed in the early segments of the response chain and slower terminal PRF speed late in the chain although there were indications of an interaction between the reward schedule and the color of the runway. In both experiments PRF and CRF responding extinguished alike. Within- and between-S groups were compared in the last two experiments. In Experiment 3 the within-S acquisition data were somewhat like those from the first 2 experiments. The between-S PRF group, however, did not show higher speeds than the CRF group early in the response chain, as is often reported, but were slower in all segments of the chain. ½ of the Ss were extinguished in 1 runway and ½ in the other making all extinction comparisons between Ss. Aside from strong effects of runway color the within-S groups showed no difference in extinction of PRF and CRF responding and extinguished more like the between-S PRF group than like the CRF group. Experiment 4 was similar to Experiment 3 except that extinction was in both runways. Acquisition and extinction findings were similar in their import to those obtained in Experiment 3. A concluding discussion points up differences between acquisition and extinction data from the within- and between-S situations and attempts to reconcile these differences in terms of frustrative nonreward mechanisms.

For several years we have been studying a class of phenomena related to successive discrimination. This is the kind of discrimination in which (as in Pavlovian differentiation) one of two discriminanda is presented on any one trial and discrimination is evidenced by changes in the intensity of performance to the positive and negative stimuli. The particular aspect of such experiments which has interested us most is the one in which the subject (*S*) is exposed to a variety of prediscrimination experiences in relation to discriminanda prior to the actual differentiation training, and the rate of discrimination learning (or resistance to discrimination) is related to these experiences. The theoretical relevance of this type of experiment and some experiments utilizing different prediscrimination procedures have been described in recent reports (Amsel, 1962; Amsel & Ward, 1965).

In one experiment, prediscrimination training prior to a black-white discrimination consisted of reinforcing the response to black on an intermittent (50%) basis, while

[1] The research described in this report was supported by grants from the National Science Foundation (GB-3772) and from the National Research Council of Canada (APT 72). We are greatly indebted to John C. Ogilvie for his advice and assistance in the statistical analysis of the data of these experiments.

[2] Now at Connecticut College.

the same response to white was being reinforced on every occasion. At first, we were interested in what effect this prediscrimination condition might have on subsequent discrimination learning, and this feature of the experiment is reported in detail in a recent monograph (Amsel & Ward, 1965). However, the focus of our interest changed from the discrimination phase of the experiment to the prediscrimination phase itself, and this is the point of departure for the present monograph. It became apparent while running the prediscrimination phase that we were involved in the study of partial reinforcement (PRF) effects within *S*s (Amsel, MacKinnon, Rashotte, & Surridge, 1964). We observed that in the prediscrimination phase the B$\pm$ stimulus was operating within a given *S* as a PRF condition, while the W+ stimulus was operating like a continuously reinforced condition, in the sense that several of the phenomena of partial-reinforcement acquisition that had previously been reported in between-*S* types of experiments (e.g., Goodrich, 1959) were also discernible within an *S* although the within-*S* and between-*S* phenomena were by no means identical.

While the phenomena uncovered by Amsel et al. seemed clear enough, the experiment itself was incomplete and deficient in several respects, the main value of the study being that it seemed to suggest an important approach to the study of mediational factors operating in partial reinforcement. It also pointed up the fact that between-*S* and within-*S* PRF conditions represented potentially important differences in psychological processes—differences it might pay us to understand better. Our present task is, then, to conduct PRF experiments that will (*a*) eliminate the deficiencies of our first within-*S* experiment, and (*b*) allow us to make meaningful comparisons between the within-*S* and between-*S* cases. We will introduce the two facets of our purpose in these experiments in reverse order.

The Study of Partial Reinforcement

The phenomena which fall under the heading of partial reinforcement effects have been among the most examined and most interpreted in psychology for the last 20 years. While this is not the place to review in detail the great variety of findings and interpretations that might be included in this area, an outline of such a review might nevertheless be in order to delimit that portion of the larger study of partial reinforcement which will concern us here.

A response is said to be partially or intermittently reinforced if it is rewarded according to some probability less than unity, and according to any of a variety of patterns. While early references to reinforcement on a schedule of less than 100% can be found in Pavlov (1927) and Skinner (1938), the major early systematic attack on the problem was by L. G. Humphreys who, from 1939 to 1943, published a variety of experiments in which he compared partial with continuous reinforcement. He studied a diversity of responses including eyelid conditioning (Humphreys, 1939a), verbal expectancy or guessing (1939b), galvanic skin response (GSR) conditioning (1940), and the acquisition of bar pressing in a Skinner box (1943), and came to the conclusion, as did an early review of the literature by Jenkins and Stanley (1950), that partial as compared with continuous reinforcement resulted in greater resistance to extinction, and that partial reinforcement acquisition was either only slightly inferior to or equal to acquisition under continuous reinforcement conditions. Since that time, several major lines of research on partial or intermittent reinforcement have been developing, and it is at least approximately the case that each of the various lines of investigation pursued by Humphreys has become a specialized area in its own right. For example, the study of verbal expectancy or guessing has been done almost exclusively by those interested in the development of mathematical models; while investigators with seemingly quite different interests have pursued the discrete-trial instrumental learning and the free-operant investigations of partial reinforcement phenomena, using mainly nonhuman *S*s.

In this report, we will be studying discrete-trial instrumental learning with moderate

spacing of trials. Therefore, our discussions will not, necessarily, have relevance for the free-operant case, nor even for the discrete-trial case with extremely short intertrial intervals. Our explanations involve classical conditioning as an explanatory model for PRF effects in instrumental behavior, but we do not investigate classical conditioning directly. An explanation which involves hypothetical, classically conditioned internal responses as mediators of overt instrumental behavior cannot, at the same time, account for partial reinforcement effects in classical conditioning, and we will take the position, as Spence (1960) has, that different explanatory schemes will probably be required to understand PRF effects in classical conditioning and instrumental learning.

It will also be clear that our phenomena are not coextensive with those studied by Skinner and his associates under the heading of *Schedules of Reinforcement* (e.g., Ferster & Skinner, 1957). The Skinnerian situation, involving free responding, is different from the discrete-trial runway situation in a variety of ways; but perhaps the most important difference is in the separation of trials, that is to say, in the integrity of the individual-trial experience. When a Skinnerian speaks of fixed-interval and variable-interval schedules, or of fixed-ratio and variable-ratio schedules, or of any of the number of other compound schedules built of these four basic elements, he is involved very importantly in the concept of response chaining: the feedback stimulation arising from the nth response may be part of the stimulus pattern affecting the $n + 1$th response; the reinforcement of the nth response may also affect the $n - 1$th response, and even other responses more remote in the chain when the organism effects a burst of responses for a terminal reinforcement. In our work, we are at some pains, as will be evident from some of our design considerations, to achieve discreteness of trials. While we may not always be entirely successful in this, the type of thinking we pursue is most effective in those situations in which the carried-over traces or aftereffects from one trial are not a part of the stimulus complex for the next trial.

Having restricted our area of interest to the discrete-trial, instrumental learning situations, we can now pursue the distinction between within-*S* and between-*S* PRF situations venturing some opinions on the psychological significance such a distinction may have. In between-*S* designs, one group of *S*s performs under conditions of partial reinforcement and the other under continuous reinforcement, and the effects of partial reward on the form of acquisition and extinction curves are examined in terms of between-group differences. In the within-*S* experiment, the same *S* is partially rewarded for a response in the presence of one stimulus (S_1) and continuously rewarded for the same response in the presence of another stimulus (S_2). The comparison in such experiments is between the performance of a group of *S*s to one stimulus as compared with the performance of the same group of *S*s to another stimulus. (Of course, this is identical in principle to the analysis of data from discrimination learning with separate presentation of stimuli.)

To reduce the comparison of the between-*S* and within-*S* cases to its most simple and ideal level would require a between-*S* experiment involving two *S*s and a within-*S* experiment involving one. (This, of course, is approximated in some of the Skinnerian reports.) In the two-*S* experiment, each of the two (presumably otherwise identical) *S*s would be exposed to a different reinforcement condition, one continuous the other partial. The one-*S* experiment would involve the development within the *S* of two separate systems or processes, one related to S_1 and the other related to S_2. It is difficult to maintain that one form of the experiment has any advantage over the other; they are simply different. The psychological justification for the between-*S* experiment would be that it represents an experimental model for how separate but similar organisms are affected by different patterns of reinforcement in relation to the same response. The within-*S* experiment, on the other hand, is a model for the development within the same organism of different systems or processes relative to different environmental events with associated rein-

forcement contingencies. An example of such between-*S* and within-*S* differences that immediately comes to mind is from that complex of stimulus-response-reinforcement relationships that exists in "the family." Forgetting for the moment about complex interactions between hereditary and environmental factors determining personality, and assuming that the family we are considering is composed of mother, father, and two children (why not identical twins?), all of the elements for our comparison are present. Let father be Stimulus_1 and mother Stimulus_2 (no priority intended); and let the children be Subject_1 and Subject_2. Each child is then an *S* in both a within-*S* and a between-*S* experiment to the extent that (*a*) each, separately, may be on a different discrete-trial schedule of reinforcement in relation to the two parents for the same behavior, and (*b*) the two children may be on different schedules of reinforcement in relation to each parent for the same kind of behavior. The complexity of the relationships that are possible, even in this next-to-simplest of family situations, will be apparent, and the kinds of question that this situation raises will be obvious. Can the same child learn different patterns of vigor-persistence relationships to the two parents as stimuli and reinforcing agents? Assuming that this is so, still, to what extent will the child who has learned persistence in relation to the inconsistent reinforcing tactics of one parent manifest persistence also in relation to the other parent who has been more consistent? These are within-*S* questions. The between-*S* questions might apply to the relationships between one parent and both children, or between the much more complex relationships that hold between both parents and both children. Between-*S* experiments, particularly those run in the laboratory and particularly, again, those using animals as *S*s, are asking very simple questions compared to those that might be asked about the family relationship. The basic question of course is, to what extent can differences in reinforcement produce two different organisms, one relatively more persistent—or more vigorous, or both—than the other?

Our introduction to between- versus within-*S* differences has been in terms of a specific example. We will develop some of these ideas in a more general and abstract way later. But for the moment we turn to the second purpose of our experiments which is to follow up our first within-*S* experiment and try to eliminate some of its deficiencies.

DEFICIENCIES IN THE AMSEL, MACKINNON, RASHOTTE, AND SURRIDGE EXPERIMENT

The experiment of Amsel et al. (1964) was incomplete in at least four important respects.

1. It was run under only one condition of color, that in which the black stimulus signaled partial and the white stimulus signaled continuous reinforcement. The clear possibility remained that some interactions of color and reinforcement pattern might be operating and that reversing the color conditions in relation to reinforcement might produce a different result.

2. The experiment was run under conditions which equated the two stimuli for numbers of reinforcements rather than trials. In the usual between-*S* PRF experiment, two groups of *S*s are run, one under continuous the other under partial reinforcement. Ordinarily, both groups experience the same number of trials during the experiment but are different with respect to numbers of reinforcements, the continuous group usually receiving twice as many as the partial group. Our experiment was run three trials a day, with two trials to the black stimulus, partially reinforced, and one trial to the white stimulus, continuously reinforced. There remained a question whether the phenomena demonstrated under the $B\pm W+$ conditions (reinforcements equated) could be replicated under $B\pm W\mp$ conditions (trials equated).

3. Our experiment contained no direct comparison of the within-*S* result with an appropriate between-*S* condition. We simply compared our result to the various between-*S* experiments that had already been published, having great confidence in the between-*S* results which had been reported in many separate and independent investigations.

4. Our report contained no extinction

data, since the experiment itself was run as the initial phase of an experiment in which the second phase was discrimination learning. The Amsel and Ward monograph to which we have referred did suggest that following many trials of B±W+ prediscrimination experience "discriminative extinction" was slower in a B−W+ discrimination than in a B+W− discrimination, providing what appeared to be evidence for differential extinctive effects attributable to within-*S* PRF acquisition; however, we had no data to tell us how the same *S* would extinguish to *both* black and white stimuli after this sort of prediscrimination training, or how a group of *S*s would extinguish to *only* black as compared to another group to *only* white when both groups had had the B±W+ preextinction experience.

THE EXPERIMENTS

The present monograph describes four experiments which (*a*) extend the Amsel, MacKinnon, Rashotte, and Surridge experiment and cover the deficiencies which have been described; (*b*) examine in a detailed fashion the kinds of acquisition and extinction phenomena that emerge out of the within-*S* PRF experiment when it is run under a variety of conditions; and (*c*) make such comparisons as are possible between the within-*S* and between-*S* experiments.

Much of what we tried to do in these experiments can be understood from a schema shown in Figure 1. This figure says, essentially, that in PRF experiments acquisition can be either between *S*s or within *S*, that extinction following within-*S* acquisition can also be either between *S*s or within *S*, and that extinction following between-*S* acquisition must necessarily be between *S*s, since there is then no basis for within-*S* extinction. In between-*S* acquisition, as shown in the top left of the schema, two *S*s or two groups of *S*s are exposed to the same stimulus conditions, but in one group the stimulus always precedes partial reinforcement, whereas in the other the same stimulus always precedes continuous reinforcement. To test for the PRE in extinction, both groups are presented with the same stimulus, but in no case is either *S* reinforced for the response it makes to that stimulus.

SCHEMA OF BETWEEN-S AND WITHIN-S
EXPERIMENTAL CONDITIONS FOR PARTIAL REWARD EXPERIMENTS

ACQUISITION		EXTINCTION	
		BETWEEN S	WITHIN S
USUAL BETWEEN S	$STIM_1 \pm$ (P)	$STIM_1 -$	
	$STIM_1$ ‡ (C)	$STIM_1 -$	
WITHIN S (PC)	$STIM_1 \pm$ (P) $STIM_2$ ‡ (C)	$STIM_1 -$ OR $STIM_2 -$	$STIM_1 -$ AND $STIM_2 -$
BETWEEN S (PP)	$STIM_1 \pm$ (P) $STIM_2 \pm$ (P)	$STIM_1 -$ OR $STIM_2 -$	$STIM_1 -$ AND $STIM_2 -$
BETWEEN S (CC)	$STIM_1$ ‡ (C) $STIM_2$ ‡ (C)	$STIM_1 -$ OR $STIM_2 -$	$STIM_1 -$ AND $STIM_2 -$

FIG. 1. The row above the double line (between-*S*) represents the usual partial reinforcement experiment; the second row represents within-*S* acquisition followed by either between-*S* or within-*S* extinction. The bottom two rows, (PP) and (CC), together constitute the appropriate between-*S* control for the within-*S* (PC) experiment.

The basic within-*S* condition (PC) is shown in the next row of the diagram. In this case the same *S* experiences two stimuli and, while response to $Stimulus_1$ is partially reinforced, response to $Stimulus_2$ is continuously reinforced. Following such within-*S* acquisition, extinction may be run in a between-*S* or within-*S* manner. In between-*S* extinction for such a condition, half of the *S*s are extinguished only to $Stimulus_1$ while the other half are extinguished only to $Stimulus_2$, all *S*s having been exposed to both stimuli during acquisition. Within-*S* extinction following within-*S* acquisition allows every *S* to be exposed to both stimuli during extinction and to experience nonreward in relation to both. In both kinds of extinction following within-*S* acquisition we are testing for the effects of reinforcement schedules applied during acquisition.

The third and fourth rows of this schema (PP) and (CC) represent our conception of the appropriate between-*S* comparison for the within-*S* experiment. These are within-*S* cases only in the sense that *S* is exposed to two stimuli as he is in the within-*S* PRF experiment. However, in the PP case *S* experiences PRF in relation to both stimuli, while in the CC case continuous reinforcement (CRF) in relation to both stimuli. Consequently, PP versus CC is the same as the between-*S* PRF at

the top of the schema, with the one difference that each S in both groups is exposed to two stimuli ($S_1\pm S_2\pm$ or $S_1\mp S_2\mp$) instead of only one, as is the case in the ordinary between-S experiment. It is in this same limited sense that such Ss may be extinguished in the between-S or within-S mode. All this means for the PP and CC cases is that an S who has had PRF in relation to two stimuli or CRF in relation to two stimuli may undergo extinction to only one of these stimuli or to both of them. To reiterate, the rows of our schema which are designated within-S (PC), between-S (PP), and between-S (CC), together constitute our conception of a within-S experiment: a group which is reinforced on two schedules (PC) is run along with appropriate between-S "control" groups (PP and CC).

There is reason to question whether the particular comparisons of between-S and within-S experimental conditions we undertook are the "correct" comparisons. How should the conditions of the two kinds of experiment be arranged in order to afford reasonable comparison? After examining this question at some length we were forced to conclude that any comparison of between-S and within-S results, in the sense that data from a control and an experimental group are compared, is probably impossible, and that all we can expect to do is to compare the *kinds* of results that can be obtained under a number of versions of the two experimental forms.

To begin with, it is obvious that the within-S experiment can be conducted in a manner that equates for number of reinforcements or for number of trials to the two stimuli. Using a generalized notation, we would designate the first case $S_1\pm S_2+$, the second $S_1\pm S_2\mp$, and we have actually conducted both forms of the experiment.

However, considering only the $S_1\pm S_2\mp$ case, the equivalent between-S experiment is difficult to arrive at. First of all, the within-S condition represents 75% reinforcement if one ignores the two colors, and 50% versus 100% reinforcement if one considers that the separate colors control separate systems within the organism. Similar considerations apply to the appropriate numbers of trials for the between-S "control." Certainly each S in the within-S condition has four trials in each $S_1\pm S_2\mp$ block, but only two trials to each of the stimuli. The question then is whether the appropriate block of trials in the between-S comparison should have four trials or two. Assuming that we choose the two-trial alternative, is the proper between-S comparison to $B\pm W\mp$ two groups, a partial $B\pm$ group and a continuous $W\mp$ group? Or is it the more usual PRF experimental condition in which color is held the same between the two groups, $B\pm$ versus $B\mp$? Or $W\pm$ versus $W\mp$? Or both? Assume, on the other hand, that one selects for the between-S comparison experiment the four-trial alternative. Now every S in the experiment will experience the same number of trials; but for the within-S group, the between-S partial group, and the between-S continuous group respectively, the overall percentages of reinforcement will be 75, 50, and 100. If this is the choice, again the same questions as to color apply for the between-S conditions. Should Ss be exposed to one stimulus color only, the same across partial and continuous groups? Or should each S see one color but a different one in the PRF than in the CRF group? Or should each S see two colors as in the within-S condition, but with the same reinforcement schedule connected to both colors ($B\pm W\pm$ versus $B\mp W\mp$)?

We selected this last alternative as representing the most reasonable way to compare between-S and within-S conditions, since it meant that each S in the between-S condition would have the same number of trials as an S in the within-S condition and would also experience two alley colors during the experiment. This produced some interesting results along with some difficulties.

Experiment 1

In the pilot experiment of Amsel et al. (1964), one group of Ss acquired a running response in a straight alley under two stimulus conditions related to two conditions of reinforcement. When the alley was black S was rewarded on half the trials, and when the alley was white S was always rewarded.

Each S ran three trials on each day, two of these black and one white, for 108 days. Measures of time to traverse the 62-in. runway were taken over the early middle and late sections of the alley: a starting measure, a running measure and a goal measure, respectively. The data show that in the first two sections of the alley Ss run faster at asymptote on partial (black) trials than on continuous (white) trials, but that in the goal region Ss run slower on partial than on continuous trials, suggesting that the usual between-group partial reinforcement acquisition effect at asymptote can be reproduced within Ss.

The purpose of the first experiment was to replicate and extend the earlier experiment. Specifically, a group was run under B±W+ conditions, as in the previous experiment, and another group was added and run with the color-reinforcement relationship reversed (W±B+). Also, extinction trials were run to determine whether the usual between-S partial reinforcement extinction effect could be reproduced under within-S conditions. The extinction mode applied here was within-S, all Ss undergoing extinction to both stimuli.

Method

Subjects. The Ss were 20 experimentally naïve male albino rats of the Wistar strain from Woodlyn Farms, Guelph, Ontario. They were about 110 days old when received in the laboratory, and about 130 days old at the start of the experiment.

Apparatus. The apparatus consisted of a pair of runways, one white the other black, each of which could be aligned with a common entry box–start box unit, painted gray.

Response-time measures could be taken over either three 1-ft. segments (starting, running, and goal) or, by extending the alley, over five 1-ft. segments, involving three intermediate running measures. Experiments 1 and 4 employed the runways in their lengthened (five-segment) form, and Experiments 2 and 3 employed the shorter version.

The entire apparatus was constructed of ¾-in. pine, and was covered throughout with ¼-in. Plexiglas, hinged to allow access to the runways. The runways were 3¾ in. high and 2⅞ in. wide. The walls and wooden floor sections were painted either flat black or flat white. The entry box (11 in. × 3 in. × 3¾ in.) and the starting section (10½ in. × 2 in. × 3¾ in.) of the apparatus were painted flat gray. The entry box was separated from the start chamber by a gray, metal, guillotine-type door. A clear ¼-in. Plexiglas door separated the start chamber from the runway portion of the apparatus.

In Experiment 1 the two straight runways, white and black, were used in their lengthened mode. Five performance measures were taken (by means of Model 120A Hunter KlocKounters activated by a photoelectric system) over five 1-ft. segments. The sequence of starting and stopping the KlocKounters was as follows: (*a*) Raising the Plexiglas start door at the entrance to the runway opened a microswitch which activated the first clock. (*b*) When S traversed 1 ft. of the runway, a photobeam was interrupted stopping Clock 1 and starting Clock 2. Clock 1, then, provided the start measure. (*c*) A second photobeam located 24 in. from the start door, when interrupted, stopped Clock 2, providing the Run I measure, and activated Clock 3, and so on, until the interruption of a photobeam located 3 in. from the end of the alley stopped Clock 5 providing a goal measure. A small metal food cup extending the width of the alley and with a lip to conceal the presence of a food pellet was suspended on the end wall of the runway, 3 in. beyond the last photobeam and 2¾ in. from the alley floor.

Metal retrace doors painted either flat black or flat white, corresponding to the walls and floors of the runways, confined S to the food cup area. The length of the goal segment was 12 in.

The apparatus was illuminated by two 15-watt bulbs suspended in milk-glass globes approximately 96 in. above the runways. The only other source of light in the experimental room was one 7½-watt frosted glass bulb suspended above the clocks.

Procedure. Three weeks before the beginning of the experiment Ss were placed on a 24-hr. food deprivation schedule. During this period each S received a daily ration of 11 gm. of Purina lab chow in its home cage with ad libitum water, and each S was handled every 2 days for a few minutes. There was no habituation to or prefeeding in the apparatus.

The Ss were run in squads of 10, with five Ss from each color group randomly assigned to each squad. Squads were taken to the experimental room in a 10-place carrying cage and waited 10 min. before the first trial of the day.

The experiment involved 270 acquisition trials and 36 extinction trials. During acquisition there were three trials per day, two to the partial stimulus and one to the continuous. The six possible trial sequences were randomly assigned within successive 6-day blocks for each S over the duration of the experiment. Ss were run with a minimum intertrial interval of 15 min.

The procedure on any individual trial was as follows: (*a*) S was removed from the carrying cage and placed in the entry box. (*b*) When S had oriented in the direction of the runway, the gray metal door was raised, allowing S to enter the start box. (*c*) Approximately 1 sec. later, the clear Plexiglas door leading to the runway was raised

and S was allowed to traverse the runway and enter the goal area where, on reward trials, the food cup was baited with one 500-mg. Noyes pellet. (*d*) The retrace door was lowered, confining S in the goal box. On a reward trial S remained in the goal box until the pellet was consumed. (*e*) S was removed from the goal box and placed back in the carrying cage. On nonreward trials the procedure was identical except that S was removed from the goal box after 15 sec. Following the last trial of a day, *S*s were placed in their home cages, and 40 min. later received the remainder of their daily ration. In order to reduce the possibility of discrimination on the basis of olfactory cues in the goal area, ground food, not of course visible to S, was scattered beneath the runway. Also, prior to each individual trial whether rewarded or nonrewarded, the goal box was swept clear of any traces of food particles from the previous trial. Consequently, the pretrial manipulations of the experimenter (E) were nondifferential over all trials, decreasing the likelihood that S might pick up a cue for reward or nonreward from E's activity. (These procedures were followed in all experiments.)

During extinction, *S*s were run as in acquisition except that food was never available in the goal box and *S*s were detained in the goal box for 15 sec. on all trials. If S failed to traverse any of the five 1-ft. segments of the alley within 60 sec., he was removed from the alley and all subsequent segments were scored as 60 sec. for that trial.

Results and Discussion[3]

The results of the first experiment are shown graphically in three figures. Figure 2 shows group curves for all alley measures and for acquisition and extinction in mean reciprocals (speeds). The acquisition data are plotted in 6-day blocks, the extinction data in 3-day blocks. Figure 3 presents individual S acquisition data for all measures for three *S*s which represent types of performance observed in this experiment. Figure 4 shows extinction performance for each S in terms of the goal measure.

Acquisition. The results as presented in Figure 2 are separated for the two color conditions. The acquisition data for the group in which black is partial (B±W+) replicate and extend the previously published data. Performance to B± is more vigorous than to W+ in all the measures beginning with the second 18-trial (6-day) block and remains so in the start and Run I measures. In the other running measures and in the goal measure this difference disappears, and in all measures it decreases, progressively and systematically working forward in time from the goal measure where a clear reversal becomes apparent. These data would tend to support an interpretation of increasing aversiveness to the partial stimulus as the goal is approached, this aversiveness becoming attached to cues earlier and earlier in the runway after extended training.

The acquisition data from the W±B+ group are shown in the left-hand panel, and they do not, at first glance, seem to show the same terminal effects, especially in the goal measure. However, close examination will reveal that the differences in performance to the two stimuli are much greater when white is the partial stimulus than when black is, and that the differences across a horizontal comparison of conditions are due largely to faster running to the partial stimulus when it is white. There seems to be little, if any, difference in speed between W+ and B+.

Analyses of variance for the acquisition data were conducted separately for three successive 30-day blocks and for each of the five measures. The three successive blocks of days are equivalent to Trial Blocks 1–5, 6–10, and 11–15 in Figure 2. Three factors were included in these analyses: a between-S factor, Color Group, and two within-S factors, Day and Reinforcement. The Reinforcement factor in these analyses refers to the three types of trial on each day (Continuous versus Partial+ versus Partial−)

[3] Analyses of variance of speed data in this report are those appropriate to split split-plot designs. In the present analyses there is a between-S factor ("Color Groups" in within-S analyses; "Reinforcement" in between-S analyses); a within-S factor ("Days") and its interaction with the between-S factor; and, a second within-S factor ("Reinforcement" in within-S analyses; "Color" in between-S analyses) and its associated interactions. Appropriate error terms were computed to test each of these effects. A complete summary table of a within- and between-S analysis is presented in Tables 1 and 2 for the fourth experiment, which is the largest in that there are five alley measures and both within- and between-S groups. This summary of the analyses typifies the treatment of speed data in all of the experiments. In the first three experiments we report only those factors in the analyses which reach acceptable confidence levels.

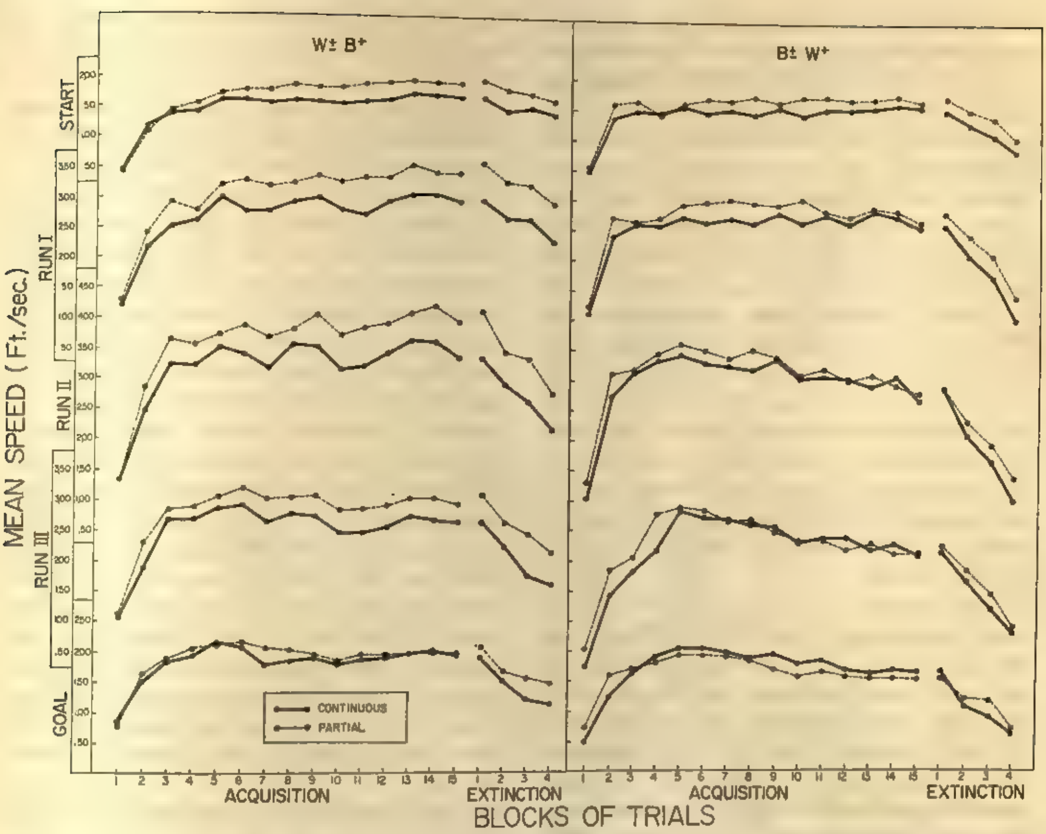

FIG. 2. Acquisition and extinction data from five alley segments for the two within-*S* color groups of Experiment 1. Acquisition data are plotted in 6-day blocks (mean of 12 partial trials and 6 continuous trials) and extinction data are plotted in 3-day blocks.

and subsequent linear contrasts allowed a partial versus continuous comparison as well as a test of the difference between the two types of partial trial. The latter test was made to determine if *S*s were discriminating between the positive and negative partial trials (presumably on the basis of firstness and secondness), for which there seemed to be some evidence earlier (Amsel et al., 1964).

Analyses for the first 30-day block show that the Day, Color Group × Day and Reinforcement effects were significant in each measure (all at the .001 level), with no other effects reaching significance. Linear contrasts on the means for each type of reinforcement showed no significant difference between the Partial+ and Partial− trials in any measure in either Color Group, and a reliable difference between partial and continuous trials only in Run II in the W±B+ group and in Run III in the B±W+ group ($p < .05$ in each case). The significant Day factor reflects the acquisition of the running response and the significant Color Group × Day interaction reflects faster running in the W±B+ group, due mainly to the higher speeds to W±.

Analyses for the second 30-day block show a significant effect of Day in all measures except Run I; of Reinforcement in all measures except goal; and of Color Group × Reinforcement in all measures except start and Run I. No other effects were significant. The Day effect reflects changes in performance across days which does not seem necessarily to be related to the acquisition of the response. Linear contrasts show no difference between the two types of partial trial in any measure in the W±B+ group, and a reliable difference of this sort only in the

start measure in the B±W+ group. Since this is the only occurrence of a difference between Partial+ and − 'trials' in the entire experiment (out of 30 such comparisons), it presumably does not reflect a systematic discrimination by the animals. Contrasts between partial and continuous trials for the W±B+ group showed significant differences in Runs I, II, and III but no significant difference in start or goal measures. For the B±W+ group there was a significant difference only in the start and Run I measures. The Color Group × Reinforcement interaction reflects the Runs II, III, and goal differences evident in Figure 2, the bulk of this difference being due to differential performance to the partial stimulus by the two groups, dropping off in the B±W+ group and remaining at asymptote in the W±B+ group.

The analyses for the last 30-day block show a significant Color Group effect in the Run III ($p < .05$) and the goal measure ($p < .01$), and this factor just fails to reach the .05 level of confidence in Run II. There is also a significant Day effect in all measures which reaches the .05 level in the start measure and the .001 level in all other measures, and a significant Color Group × Day effect in Run II and Run III ($p < .01$ each) and in goal ($p < .05$). These effects reflect an overall lower level of performance by the B±W+ group in the last three alley measures and an overall change in performance across days in this block of trials which takes the form of a decrease in running speed to both stimuli in the B±W+ group and even a slight increase in running speed by the W±B+ group. Reinforcement was also significant in all measures in this block of trials and Color Group × Reinforcement was significant in all measures except start. Linear contrasts revealed no difference between the two partial trials in either group, but a reliable difference between partial and continuous trials in the W±B+ group in all measures except goal. There was no reliable difference between partial and continuous trials in any measure in the B±W+ group. The bulk of the Color Group × Reinforcement interaction would again seem to be due to differential performance to B± and W±, the former continuing the drop off begun in the previous block of trials.

The statistical analyses of the acquisition data support the relationships apparent in the plot of speed data in Figure 2.

In Figure 2, the failure of reversal in the goal measure when the partial stimulus is white could be a purely physical artifact of the greater speed to the W± stimulus. The direction of change in the relationship of the partial and continuous stimuli from start to goal, that is, the successive slowing down of performance to the partial *relative to the continuous stimulus*, seems to be the same in both cases; the W± group, however, starts out faster than the B± group. It is as if the goal panel on the left side of Figure 2 belongs at about the level of the Run II or Run III panel on the right. This color difference seems to appear in some of our experiments but not others, and it is obviously a stimulus intensity effect characteristic of within-*S* designs in which adaptation level (Helson, 1964) and contrast factors would naturally be important (Beck, 1963; Grice & Hunter, 1964).

A graphical plot of individual-*S* acquisition data suggests that all *S*s show the progressive slowing down of performance to the partial relative to the continuous stimulus as the goal is approached but that there are three main ways in which this relative slowing down manifests itself. Figure 3 shows individual acquisition data for *S* 11, from the W±B+ group, and *S*s 10 and 4 from the B±W+ group *S* 11 is representative of 7 of the 20 *S*s, 6 from Group W±B+, which show progressive slowing down to the partial stimulus but no cross over of the curves, even in the goal measure. *S* 10 is representative of six of the *S*s, all from Group B±W+, which show faster running to the partial stimulus in the early measures and a progressive slowing down to both stimuli as the goal is approached, the amount of slowing being greater to the partial stimulus. Slower running to the partial than to the continuous stimulus, which is characteristic of the goal measure in between-*S* experiments, is found as early as the first running measure in this type of animal.

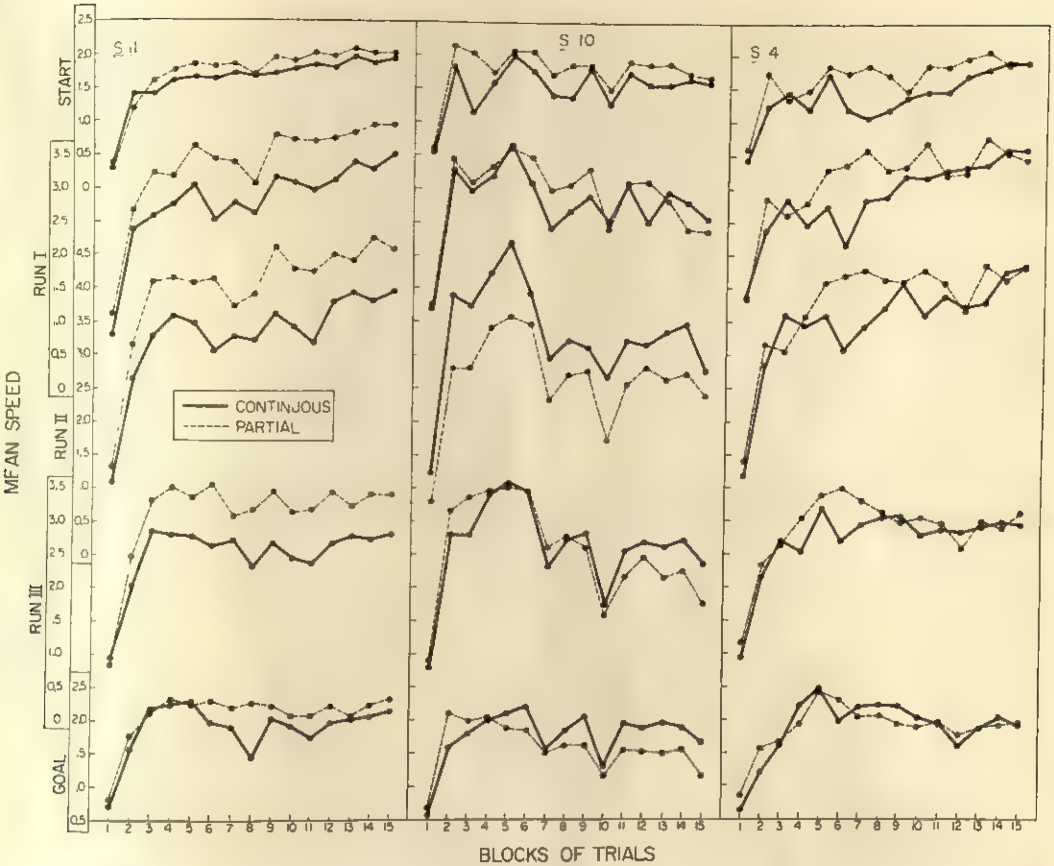

FIG. 3. Data for three individual *S*s from Experiment 1 representing three patterns of responding to $S_1\pm$ and S_2+ in the five measures. The data are plotted in 6-day blocks.

Finally, *S* 4, which is representative of six *S*s, four from Group W±B+ and two from Group B±W+, shows faster running to the partial stimulus in the start and running measures and a tendency toward slower running to the partial stimulus in the goal. This type of *S* most closely resembles between-*S* findings (Goodrich, 1959) and the within-*S* findings of Amsel et al. (1964). One *S* in the B±W+ group showed no difference between partial and continuous performance in any measure except the goal, where responding was slower to the partial stimulus.

These findings suggest that PRF and CRF, manipulated within *S*s, influence vigor of responding in a consistent fashion: there is, in almost every case, relatively more slowing to stimuli signaling PRF as the goal is approached. However, magnitude and direction of absolute differences between PRF and CRF acquisition performance over segments of the runway vary among individual *S*s. These individual differences in reaction seem clearly to be related in part to the physical intensity of the partial stimulus (white or black); they may also be related to differences in "tolerance" for aversiveness (anticipatory frustration) among *S*s, that is to say, to differences in "personality." Davenport (1963a) has reported a related finding in an experiment in which magnitude rather than percentage of reward to S_1 and S_2 was manipulated. This was basically an experiment on spatial discrimination (choice); however, forced trials were also administered to both stimuli, one associated with a five-pellet reward the other with a one-pellet reward, and on these trials all *S*s showed relatively similar patterns of

starting speed over trials to the five-pellet stimulus, but four different response patterns to the one-pellet stimulus.

Extinction. The extinction data for each measure and condition are presented to the right of the corresponding acquisition curves in Figure 2. In this experiment, a within-*S* PRE does seem to occur, at first glance. However, closer inspection suggests that the extinction differences (*a*) reflect mainly acquisition differences and (*b*) do not represent the true differences in slope characteristic of differences in persistence between partial and continuous groups in between-*S* experiments.

The analyses of variance for the extinction block of trials show a significant effect of Color Group at the .01 level in all measures except start which is at the .05 level; a significant effect of Day in all measures ($p < .001$); and a significant Color Group × Day interaction in all measures except Run II ($p < .01$ for start and $< .001$ for the other measures). The Color Group × Day effect reflects faster overall extinction in the B±W+ group. Reinforcement is significant in all measures ($p < .001$), and Color Group × Reinforcement is significant only in the Run II and III measures ($p < .05$ and .01, respectively). These two effects reflect terminal acquisition levels of performance, and they, along with the lack of a Reinforcement × Day effect, indicate that the differences in extinction in Figure 2 are due to acquisition asymptote and are not the result of different rates of extinction to the

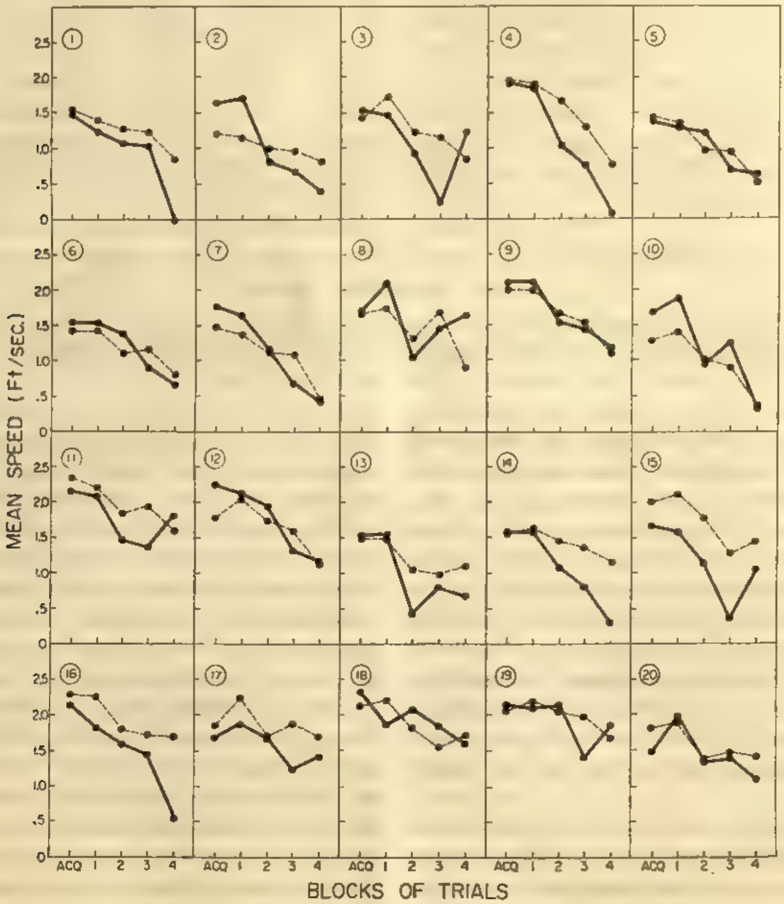

Fig. 4. Individual animal extinction data for the goal measure for 20 *S*s following $S_1 \pm S_2 +$ acquisition. The data are plotted in 3-day blocks. The solid line represents performance to S_2+, the dotted line to $S_1\pm$.

partial and continuous stimuli, which is the indicant of the PRE. Linear contrasts showed reliable differences between continuous and partial extinction performance in all measures except goal for the W±B+ group, and no differences in any measure for the B±W+ group, thereby verifying the graphical picture in Figure 2.

It is possible to have a closer look at this kind of thing in the case of within-*S* experiments by looking at the performance of individual *S*s to both stimuli. Figure 4 shows individual performance curves for the goal measure for all 20 *S*s. *S*s 1–10 are in the B± group, and *S*s 11–20 are in the W± group. There is some suggestion that *S*s 1, 2, 3, 4, and 10 show PRE-like differences in slope in the goal measure as do *S*s 13, 14, and 16. In no case was there a suggestion of faster extinction to the partial stimulus. Although the overall analysis provides no basis for a PRE in this experiment, these individual-*S* data seem to suggest that differences in rates of extinction can be observed in within-*S* PRF experiments. However some caution must be exercised: the data presented were gathered under procedures in which the number of *rewards* (not *trials*) to the PRF and CRF stimuli were equated; the more usual procedure in between-*S* experiments is to equate number of trials. In between-*S* studies involving as many trials and as large a reward as we employ, giving twice as many rewarded trials to the continuous group would facilitate extinction. Arguing from this, our procedure should not particularly favor the appearance of a PRE. The procedure of running twice as many extinction trials per day to the previously partial stimulus, as we did in this experiment, should also be unfavorable to the appearance of a PRE.

Experiment 2

Experiment 2, like Experiment 1, was a within-*S* PRF procedure followed by within-*S* extinction. There were a number of minor differences between the two experiments and one major difference: while the first experiment was conducted with reinforcements equated to the two stimuli ($S_1 \pm S_2 +$), the second equated for trials ($S_1 \pm S_2 \mp$), the more usual case in between-*S* experiments. The other difference of importance was that Experiment 2 was conducted in the shorter version of the apparatus already described, involving only three runway measures instead of five.

Method

Subjects. The *S*s were 10 male hooded rats from the colony maintained by the Department of Psychology, University of Toronto. They were 115 days old at the beginning of the experiment.

Procedure. Twenty-five days prior to the beginning of experimental training, *S*s were housed in individual cages and put on a 23-hr. deprivation schedule. They were fed 10 gm. Purina lab chow daily, and water was available at all times throughout the experiment. During the establishment of the deprivation schedule, *S*s were removed from their cages and handled by *E* for 5 min. each day. No other special pretraining procedures or habituation to the apparatus were carried out.

The *S*s were run four trials a day in the shorter version of the apparatus. Five *S*s ran under the B±W∓ condition while the other five *S*s ran B∓W±. On any experimental day, each *S* ran two trials to each stimulus. The 12 possible orders of trials were randomized, separately for each *S*, within each 12-day block.

Training was carried out over a period of 42 days (84 trials to each stimulus). Within any single trial, the procedure followed was identical to that in Experiment 1. The intertrial interval was about 15 min., and the reward magnitude, where applicable, was 500 mg. given in one Noyes pellet.

Extinction, also, was conducted on a four-trial-a-day basis, again with an intertrial interval of 15 min.

Following the final daily trial, *S*s were transported back to their home cages where they received the balance of their 10-gm. daily ration 40 min. later.

Results

The results of the second experiment are presented in Figures 5 and 6. Figure 5 presents all of the data in 2-day blocks (four trials to each stimulus) for acquisition and extinction, averaged for all *S*s across the two color groups. The most obvious features of these data, as compared to the data of Experiment 1 (see Figure 2), are (*a*) that the acquisition differences evident in the start measure in Experiment 1 were not evident in the start measure of Experiment 2; (*b*) that the greater running speed to the partial stimulus which was a characteristic of all three running measures in Experiment 1

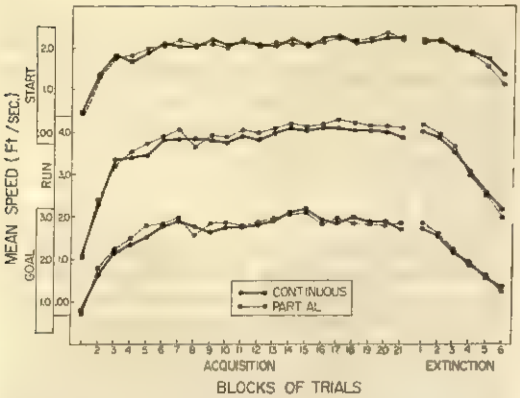

FIG. 5. Within-S acquisition and extinction for combined color conditions of Experiment 2. Acquisition and extinction are plotted in 2-day blocks of trials (mean of four partial and four continuous trials).

was also clearly evident in the one running measure of Experiment 2; (*c*) that the appearance, early in training, of greater speeds to the partial stimulus than to the continuous in the goal measure, suggested in Experiment 1, was also clearly evident in Experiment 2; (*d*) that the ultimately slower partial than continuous speeds found in the goal measure in Experiment 1, at least in the B±W+ condition, was not evident in the combined data of the present experiment (and was not found in either of the color conditions); and (*e*) that there seemed to be, again, little, if any, difference in the rates of extinction to the partial and continuous stimuli.

Separate analyses of variance were conducted for each of the three measures for two segments of acquisition (Days 1–20 and Days 21–42) and for extinction. The factors in these analyses were Color Group (a between-S factor), and Day and Reinforcement (within-S factors), the Reinforcement factor testing continuous against partial speeds.

The analyses show that the Day factor is significant at the .01 level in all measures throughout the entire experiment, which points up the fact that the learning curves are changing even beyond the first 20 days (80 trials). Perhaps the most noteworthy aspects of the analysis are those related to the Reinforcement factor. Reinforcement and interactions involving reinforcement are not significant in any segment of the experiment in either the start or goal measures. In the first 20-day segment of the run measure, Reinforcement and Reinforcement × Day are significant ($p < .05$). Over the latter 20-day segment of acquisition, Reinforcement (but not the Reinforcement × Day interaction) continues significant in the running measure ($p < .01$). All of this is statistical corroboration of characteristics of the graphical analysis that have already been described.

A further point to be noted is that the analysis of extinction shows no Reinforcement effects and, in particular, no significant Day × Reinforcement interactions, which would, if they had been present, indicate a within-S PRE.

The only other significant effects occur in relation to color. During acquisition there is a significant Color Group effect in the goal measure during the first 20 days ($p <$ .05), and in the start measure during the last half of acquisition ($p < .05$). This Color Group effect simply reflects a difference in absolute speed: the group run under B±W∓ conditions tends to start faster than the group run under W±B∓ conditions, and a difference in the same direction occurs in the goal measure during the first 20 acquisition days. In extinction there is a Day × Color Group interaction in all measures, as well as a Color Group effect in the start measure, which indicates that the W±B∓ group extinguishes more rapidly (to both stimuli) than the B±W∓ group.

In Figure 6 we present extinction data for each of the 10 Ss of the experiment, separately, for each of the three performance measures. Each panel of this figure portrays extinction data for a single S for a single measure to the two stimuli to which the S reacted in extinction, one associated with partial reinforcement the other with continuous. If there is any pattern to these data within-Ss, it is a suggestion, not borne out by the statistical analysis, that Ss extinguish faster to white than to black regardless of the pattern of reinforcement that has been associated with these stimuli during acquisition. The one fact that does seem to stand

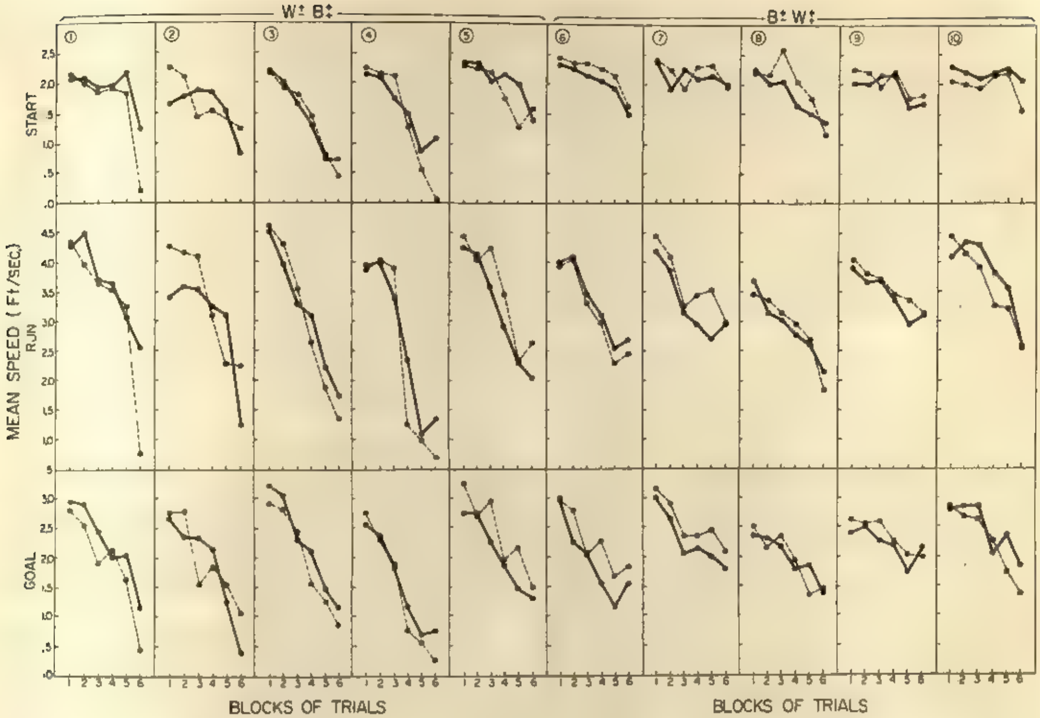

FIG. 6. Individual *S* extinction data for all 10 *S*s and for all three runway measures in Experiment 2. The solid curve represents performance to $S_2\mp$, the dotted curve to $S_1\pm$. Data are plotted in 2-day blocks of trials.

out very clearly is that there is an obvious difference between the color groups, extinction slopes to both stimuli being generally much steeper for W±B∓ than for B±W∓. This was, of course, supported by the significant Day × Color Group interaction of the statistical analysis. One could argue from this pattern of results, and from the lack of a Reinforcement × Day interaction, that for each *S* the rate of extinction to the two stimuli is not determined by the schedule of reinforcement associated with each but by the color of the partial stimulus, being greater for W± than for B±. The conjecture then would be that the characteristics of the extinction response to the partially reinforced color generalized to the continuously reinforced color within the same *S*s, and that the Day × Color Group interaction which shows up as significant in two measures is mainly due to the color of the partial stimulus. Another way of looking at this is that the white stimulus, relative to the black, is more effective in generating whatever the dominant tendency is at the moment; that in acquisition the white stimulus generates more "excitement," while in extinction it generates more "inhibition."

INTRODUCTION TO EXPERIMENTS 3 AND 4

The first two experiments involved a within-*S* acquisition procedure followed by within-*S* extinction. In Experiment 1 reinforcements were equated between S_1 and S_2, whereas in Experiment 2 trials were equated. The procedure of equating for reinforcements, which was also the procedure of the pilot experiment (Amsel et al., 1964), provided a replication of the pattern of acquisition effects in the original experiment when the stimulus conditions were the same (B±W+). There appeared in this experiment to be some suggestion of a within-*S* PRE. However, these were not as dramatic as those ordinarily seen in between-*S* experiments; were not significant in group analysis; and were, at most, merely suggested by an examination of the data from individual *S*s.

The second experiment, in which trials were equated, also provided a clear picture of more vigorous performance to the partial than to the continuous stimulus, but only in the run measure—neither the start nor the goal measure showed any evidence of differences of performance to the two stimuli. The extinction curves to $S_1\pm$ and $S_2\mp$ were virtually identical, and even an examination of the individual animal data revealed no evidence of differences in extinction to the two stimuli.

Experiments 3 and 4 were conducted for the purpose of looking further into within-*S* PRF effects under conditions of equated trials to S_1 and S_2 (rather than equated reinforcements). Each of these experiments was of the $S_1\pm S_2\mp$ type and was conducted with a four-trial-a-day procedure. Experiment 3 was run in the shorter version of the apparatus involving three runway measurements, while Experiment 4 was run in the longer, five-measurement runway. In Experiment 3, *S*s were extinguished under a between-*S* procedure following within-*S* acquisition (see Figure 1), that is to say, half of the *S*s were extinguished only to S_1 while the other half were extinguished only to S_2. In Experiment 4 the extinction procedure was, as in the first two experiments, within-*S*, all *S*s being extinguished to both S_1 and S_2.

The particular importance of Experiments 3 and 4 is, however, that in both experiments between-*S* conditions were run in parallel to the within-*S* condition, so that some comparisons of the results of the two procedures could be made.

Experiment 3

Method

Subjects. The *S*s were 40 experimentally naïve, male albino rats, obtained from Woodlyn Farms, Guelph, Ontario. Their age at the beginning of the experiment proper was approximately 90 days. Two *S*s died during the course of the experiment, leaving 38 *S*s from which data were collected.

Apparatus. The apparatus was the same as that described in Experiment 2.

Procedure. Upon arrival at the laboratory, all *S*s were assigned to individual living cages and placed on an ad libitum diet of food and water for 2 days. Fifteen days prior to the beginning of acquisition training, *S*s were placed on a 23-hr. food deprivation schedule. During this period each *S* received a daily ration of 10 gm. of Purina lab chow in its home cage. Water was available at all times. During this 15-day adjustment phase of the experiment, each *S* received four 5-min. gentling sessions. No other habituation or prefeeding procedures were carried out.

The *S*s were run in squads of 10, with animals from all experimental groups randomly assigned to each squad. The *S*s were taken to the experimental room in a 10-compartment carrying cage, and waited for a period of 10 min. before the first trial of the day. Each *S* received two trials on Days 1 and 2, and four trials per day for the remainder of the experiment (152 acquisition trials in 39 consecutive days). During acquisition the *S*s in any squad were run in a different order each day, and the minimum intertrial interval was 15 min.

The procedure on individual trials was exactly as described earlier, as were the manipulations of the experimenter before each trial which were designed to insure that olfactory and auditory cues, and visual cues other than those provided by the runways, were approximately constant from trial to trial and not differentially related to reward and nonreward.

The *S*s were randomly assigned to three experimental groups, corresponding to the designations CP, PP, and CC of Figure 1. In terms of our generalized notation, Group CP becomes $S_1\pm S_2\mp$; Group PP, $S_1\pm S_2\pm$; and Group CC, $S_1S\mp S_2\mp$. The first is the within-*S* group, and the second and third are the "controls" we selected to run, a between-*S* partial group and a between-*S* continuous group with within-*S* exposure to two stimulus colors.

Group $S_1\pm S_2\mp$ corresponds to the condition run as Experiment 2: running to one runway brightness ($S_1\pm$) was partially reinforced, and running to the other runway brightness ($S_2\mp$) was continuously reinforced. For 10 of the 19 *S*s in this (within-*S*) group, $S_1\pm$ was the white alley, and $S_2\mp$ was the black; for the other 9 *S*s these relationships were reversed. On any experimental day, each *S* of $S_1\pm S_2\mp$ ran 2 trials to $S_1\pm$ and 2 trials to $S_2\mp$. The 12 possible orders of daily presentation of trials were randomly assigned to successive 12-day blocks of training for *S*s under one color condition (e.g., $B\pm W\mp$). *S*s running to the other stimulus arrangement received the reverse order of trials with respect to color. (For instance, if the *S*s running to $B\pm$ ran $B-W+W+B+$ on a daily block of trials, the *S*s running to $W\pm$ ran $W-B+B+W+$).

The *S*s in Group $S_1\pm S_2\pm$ were rewarded 50% of the time to both the black and white stimuli, while *S*s in the third Group ($S_1\mp S_2\mp$) were continuously reinforced for approach to both stimuli. These two groups together constituted a between-*S* experiment run under conditions as similar as possible to the within-*S* procedure. (See our discussion of this problem in the introductory comments.)

Thirty-six extinction trials were run in 9 con-

secutive days. The *S*s were extinguished to one stimulus only. Half of the *S*s in each of the three groups were extinguished in the S_1 alley and half in the S_2 alley.

Results

In this experiment, as in Experiment 4, we will be comparing data collected under within-*S* and between-*S* conditions. Figure 7 shows the results of this experiment for all measures, both in acquisition and in extinction, and for between-*S* and within-*S* conditions. The acquisition data are plotted in 2-day blocks and the extinction data in daily blocks of trials. It is important to remember that in Figure 7 each panel represents, for one measure, one color version of the within-*S* experiment (e.g., $B\pm W\mp$) with its associated between-*S* counterpart. The between-*S* experiment was described in the introduction as CC versus PP. CC refers to the fact that a group of *S*s is run under *continuous reinforcement* conditions to two separate stimuli ($B\mp W\mp$); PP is the condition in which the same *S*s are run under *partial reinforcement* conditions to two stimuli ($B\pm W\pm$). In each panel, therefore, we compare our within-*S* condition ($W\mp B\pm$) to *S*s with the same reinforcement-color combinations from the between-*S* condition (e.g., $W\mp$ from the CC group and $B\pm$ from the PP group). Likewise, the $W\pm B\mp$ within-*S* condition, in the left-hand panels of Figure 7, is compared with responses to $B\mp$ from the CC condition and $W\pm$ from the PP condition. These kinds of arrangements (which hold in this experiment and the next) seem to us as good as can be made for comparing data from within-*S* and between-*S* PRF experiments.

As we look at these data we should remember that, in this experiment, extinction was between-*S* in all cases; that is to say, each extinction curve represents a separate group of *S*s extinguished to only one color. In all cases, every *S* saw both colors during acquisition. Therefore, the only factor that differentiated *S*s during extinction was whether their acquisition had been within-*S* or between-*S*. It boils down to this question: Will *S*s who acquire a response under $W\pm B\mp$ conditions and are then split in extinction, so that half of them extinguish to

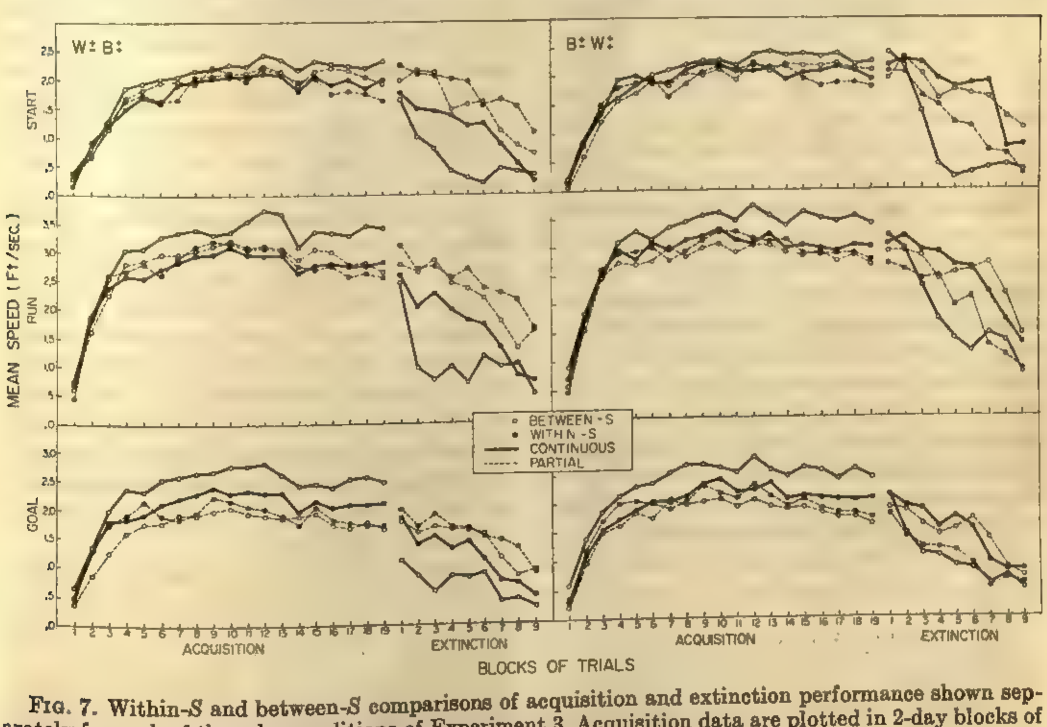

FIG. 7. Within-*S* and between-*S* comparisons of acquisition and extinction performance shown separately for each of the color conditions of Experiment 3. Acquisition data are plotted in 2-day blocks of trials, extinction data are plotted in daily blocks.

W and half to B, show the same extinction differences as might appear between a group extinguished to B after W∓B∓ acquisition and a group extinguished to W after W±B± acquisition?

Extinction. Looking first at the extinction data, it is clear that the question we just posed must be answered in the negative. Comparing the $S_1 \pm S_2 \pm$ and $S_1 \mp S_2 \mp$ groups (open circles) there is clear evidence of the typical PRE in every panel of the graph even though all *S*s were exposed to two stimuli during acquisition rather than (the usual) one and were then extinguished under only one of the two stimuli. What seems fairly clear from these data is not only a difference in slope but also a difference in acceleration (and this may turn out to be the most unmistakable indicator of the genuine PRE): the continuous group in extinction shows a strong negative acceleration, whereas in the partial groups acceleration seems to be, if anything, positive.

Between-*S* extinction following within-*S* PRF training seems not to show these effects, although at first glance there appear to be differences between the extinction groups. The left-hand panels show the partial-stimulus group apparently extinguishing more slowly than the continuous-stimulus group whereas the right-hand panels seem to show the reverse. However a closer examination shows that there are no slope differences between these extinction groups following within-*S* acquisition, and that, in fact, the slopes of both groups seem very similar to the slope of the comparison between-*S* partial group in extinction. Further examination of these extinction data will show that the apparent differences in between-*S* extinction that show up in the groups after within-*S* acquisition are largely, if not entirely, due to color, the case being that *S*s extinguished to white show generally higher performance levels in extinction than the *S*s extinguished to black. This shows up in the statistical analysis[4] as a significant Color effect in the start ($p < .01$) and run ($p < .05$) measures and an almost significant effect in the goal measure ($p < .10$). There are no significant interactions between Color and Day or between Color and Reinforcement in any analysis. This would suggest that the observed differences between within-*S* partial and continuous curves in extinction are due to color alone, and that in the absence of such color effects the two within-*S* extinction curves are almost inseparable from one another and from the between-*S* partial extinction curve. The combined curves are shown as Figure 8, and the extinction curves strongly support these conclusions.

On the other hand, analyses of variance of the data from extinction following between-*S* acquisition show significant Reinforcement effects in all measures ($p < .001$, .01 and .05 for start, run, and goal measures respectively), as well as significant Reinforcement × Day interactions in all measures ($p < .001$ in start and run and $< .05$ in goal), this latter reflecting differences in slope and being the critical indicant of the PRE. There are also in the between-*S* extinction data significant Reinforcement × Color ($p < .05$) and Color × Day ($p < .01$) interactions in the running measure; and, a significant triple interaction among Reinforcement, Color, and Day in the goal measure ($p < .05$).

What all of this suggests is that (*a*) the usual PRE shows up very clearly following between-*S* acquisition but that there is no indication of a PRE following within-*S* acquisition; and (*b*) color is a factor in both the within- and between-*S* groups in extinction but affects each differently. While color interacts with both Reinforcement and Day in between-*S* extinction following between-*S* acquisition, it does not interact with any factors in between-*S* extinction following within-*S* acquisition.

Acquisition. The acquisition data of Figure 7 present a rather complex picture, and we will be interested in the between-*S* effects,

[4] The analyses of variance for extinction in this experiment are all between-group analyses since in the within-*S* groups half of the animals are extinguished to each color. As a consequence, the between-*S* factors in both within- and between-*S* extinction analyses are Reinforcement, Color, and Reinforcement × Color. The within-*S* factors are Day, the interactions Day × Reinforcement and Day × Color, and the triple interaction.

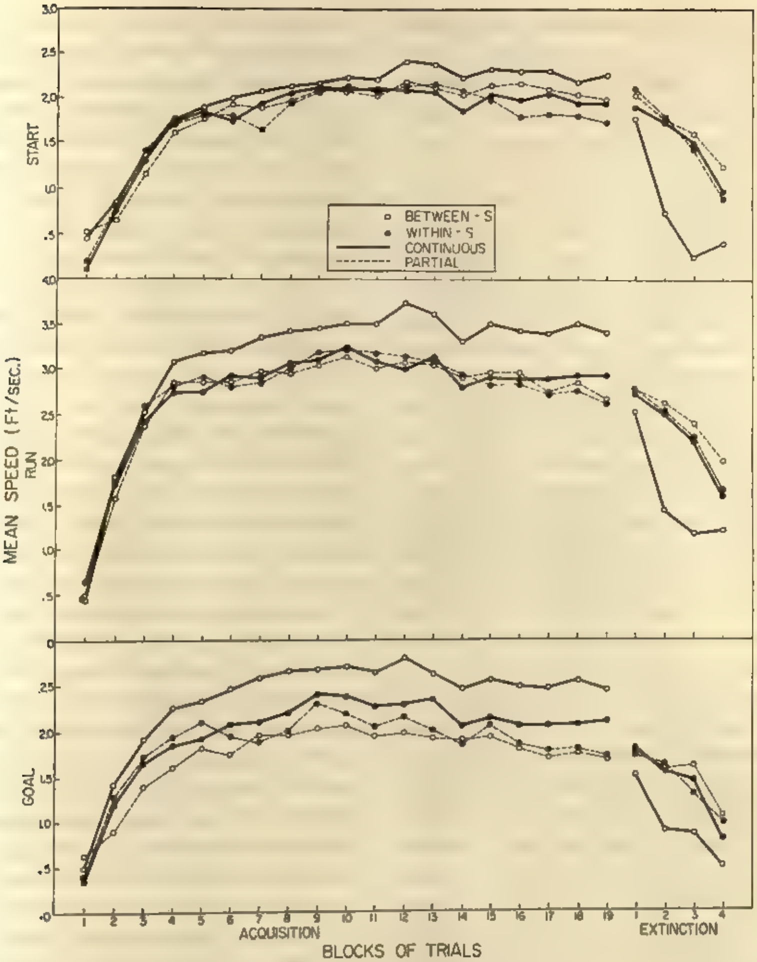

FIG. 8. Within- and between-*S* acquisition and extinction data from Experiment 3 combined across color conditions.

the within-*S* effects, and a comparison of the two.

Between-*S* effects can be summarized rather quickly. In every measure for both color groups there appears to be a difference in speeds in all cases in favor of the continuous stimulus. This is not the ordinary partial reinforcement acquisition effect as found in experiments of Goodrich (1959), Wagner (1961), and others. While we do find a progressive decrease in this difference as we go from the goal to the start measure, until at the start the difference is very small, there is no evidence of the actual crossover effect in acquisition that sometimes shows in the starting and running measures. The acquisition data are plotted in 19 2-day blocks, and for purposes of statistical analysis the entire period was divided in half, ignoring the first block of trials. A separate statistical analysis was made for each measure and for the two halves of acquisition (Days 3–20 and Days 21–38) which correspond to Blocks 2–10 and 11–19.

The analyses of variance for acquisition were done separately for within- and between-*S* data. In the between-*S* analyses Reinforcement is a between-*S* factor, and Day and Color are within-*S* factors; in the within-*S* analyses Color Group is the between-*S* factor, and Day and Reinforcement are within-*S* factors.

Analyses of the between-S acquisition data showed a very clear picture. In all measures and for both segments the Day effect was significant. In the start measure the Reinforcement effect, partial versus continuous, was not significant from Days 3–20 but was significant at the .05 level for Days 21–38. The same pattern of significance for the Reinforcement factor was found in the run measure: it was not significant over the first segment but was significant at the .05 level for the second. In the goal measure, however, Reinforcement was significant beyond the .01 level in both the early and the late segments of acquisition. The only other significant effects in the between-S data were Color effects in the first segment of acquisition in the run measure ($p < .01$), a significant Color effect in the first segment of the goal measure ($p < .01$), and a significant Color $\times$ Day interaction ($p < .01$) in the first segment of acquisition in the run measure, and in the second segment in the goal measure ($p < .05$).

The within-S curves of Figure 7 (solid dots) present an acquisition picture which is not too dissimilar to the between-S picture at the end of acquisition but which differs considerably early in training. By the end of acquisition both groups show, in all measures, faster responding to $S_2\mp$ than to $S_1\pm$, but these differences are considerably smaller than those observed in the between-S curves. There is, then, at the end of acquisition, no evidence of the partial reinforcement acquisition effect (faster to $S_1\pm$) observed in the early alley measures in Experiments 1 and 2, but a clear goal measure effect is in evidence. Earlier in acquisition there is some evidence for this effect in the running measure when white is the $S_1\pm$, and faster speeds to the partial stimulus are clearly seen in the goal measure early in acquisition, that is, up to about the fifth or sixth acquisition block. In the start and run measures both within-S curves follow, roughly, the contour and level of the between-S partial curve, and in the goal measure the within-S curves begin and remain approximately intermediate between the between-S curves.

Analyses of variance were conducted for each measure and over the same two segments of acquisition as were analyzed for the between-S groups. In each measure and in both the first and second halves of acquisition there is a significant Day effect ($p < .001$ in all cases). In addition to this overall effect a pattern of significance related to Reinforcement emerges. In the start measure Reinforcement is significant in the first segment of acquisition ($p < .05$) and just fails to reach significance in the second segment, reflecting an overall tendency for starting to the partial stimulus to be slower than to the continuous stimulus. A significant Reinforcement $\times$ Day interaction ($p < .001$) in the second segment of acquisition (Blocks 11–19) reflects faster starting to the partial stimulus beginning about Block 10 followed by relatively marked slowing about Block 15 which persists until the end of acquisition. In the run measure, Reinforcement fails to reach significance as a main effect in either segment of acquisition but does appear as a Color Group $\times$ Reinforcement interaction ($p < .01$) in the first half of acquisition and a Day $\times$ Reinforcement interaction ($p < .001$) in the latter part of acquisition. The first interaction reflects the tendency for running to $S_1\pm$ to be (*a*) faster than to $S_2\mp$ for the W$\pm$B$\mp$ group and (*b*) slower than to the $S_2\mp$ for the B$\pm$W$\mp$ group during the first half of acquisition, again suggesting that the white stimulus seems to energize the dominant response tendency. The second interaction reflects the more rapid slowing to $S_1\pm$ as opposed to $S_2\mp$ as training progresses, though performance to both tends to grow weaker as training progresses. This finding was also evident in the start measure. In the goal measure the main effect of Reinforcement was not significant in the first segment of acquisition, but a Color Group $\times$ Reinforcement interaction ($p < .05$) and a Day $\times$ Reinforcement interaction ($p < .01$) reflect the clearer difference between partial and continuous curves very early in training in the B$\pm$W$\mp$ group. In the second half of acquisition, however, Reinforcement, and Day $\times$ Reinforcement are significant ($p < .001$, and .05, respectively) reflecting the differences between groups and across days seen in Blocks 11–19 in Figure 7. The main

difference between the color groups again appears to be relatively slower running to W± than to B± suggesting that the inhibitory effect near the goal in within-*S* acquisition is greater and sets in earlier when white is partial.

Attention should be called to the difference between the kind of result obtained in within-*S* training and that obtained in between-*S* training in the earliest blocks of the goal measure. In between-*S* training, performance for the continuous group is superior from the very outset to that of the partial group. In within-*S* training, where the same *S*s run to both partial and continuous reward the picture is a rather different one. (This same effect is apparent in Figures 2 and 5 of Experiments 1 and 2, respectively; however, the between-within comparison of Experiment 3 makes it stand out even more clearly.) At a relatively early stage of training, *S*s run faster to $S_1\pm$ than to $S_2\mp$; then, between the fifth and sixth blocks of trials, the curves cross and the partial curve remains below the continuous curve until acquisition training is terminated. This relationship can be seen somewhat more clearly in Figure 8 where the color curves are combined. Estimates of the slope for individual *S*s of the within-*S* group over the first five blocks of acquisition trials were obtained, and a subsequent test of the differences in slope to $S_1\pm$ and $S_2\mp$ approached significance.

EXPERIMENT 4[5]

Experiment 4 is a combination of a within- and a between-*S* experiment conducted at separate times and is, in some respects, like the previous three but in some respects different from each of them. Like Experiment 3 it contains both within-*S* and between-*S* conditions, but unlike Experiment 3 (*a*) it is run in the longer (five-segment) runway to get a better look at transitional effects from one segment to the next in acquisition, and (*b*) it involves a within-*S* rather than a between-*S* extinction procedure. In these latter respects, then, the present experiment is like Experiment 1; but unlike Experiment 1, trials are equated rather than reinforcements in the within-*S* condition and a between-*S* comparison condition is added.

Method

Subjects. The *S*s in each experiment, within- and between-*S*, were 20 experimentally naïve male albino rats, about 110 days old at the beginning of experimental training. During the course of training one *S* died in the within-*S* experiment and one *S* was discarded in the between-*S* experiment because of illness, leaving 19 *S*s from which data were collected in each experiment.

Apparatus. The apparatus was the same as that used in Experiment 1.

Procedure. Three weeks before the beginning of each experiment *S*s were placed on a 23-hr. food deprivation schedule. During this period each *S* received a daily ration of 11 gm. of Purina lab chow in its home cage. Water was available at all times. During this 3-week period each *S* was handled every 2 days for a few minutes. No habituation or prefeeding procedures were carried out.

In each experiment the original random assignment of *S*s was into two groups of 10 each. As in Experiment 3, one within-*S* group ran under B±W∓ and the other under W±B∓ conditions, while one between-*S* group ran under B∓W∓ and the other under B±W± conditions.

In each experiment *S*s were run in two squads of 10, with 5 *S*s from the B±W∓ and W±B∓ groups in each of the two squads in the within-*S* experiment, and 5 *S*s from the B±W± and B∓W∓ groups in each of the two squads in the between-*S* experiment. Squads were transported to the experimental room in a 10-place carrying cage and waited 10 min. before the first trial of the day.

A total of 160 acquisition trials were run in each experiment, followed by 48 extinction trials in the within-*S* experiment, and 64 extinction trials in the between-*S* experiment. In both experiments there were four trials on each day, two to S_1 and two to S_2. Speed data are presented from the between-*S* experiment for only the first 48 extinction trials, since no significant speed differences occurred after this point. The additional trials were for the purpose of gathering retrace data which are a subject of analysis in this experiment.

In the within-*S* experiment the order of presentation of stimuli was determined by randomly assigning six permutations of four trials (B+, B−, W+, W+) to days within successive 6-day blocks over the duration of the experiment. This randomization was done separately for each *S*. The six orders of stimulus presentation were obtained by arranging the trials with the restriction that the first and second trials of each day be to different stimuli. In the between-*S* experiment

---

[5] We are indebted to Bohdan P. Kolesnik who assisted in the collection of data in this experiment.

the same order of presentation of stimuli was employed; but for the B∓W∓ group all trials were rewarded, while for the B±W± group only one trial to each stimulus was rewarded on each day. The *S*s were run with a minimum intertrial interval of 15 min. The details of procedure on any individual trial were exactly as described in the first three experiments.

One *S* in the W±B∓ group of the within-*S* experiment died during the course of training. For purposes of statistical analyses the *S* with the same number in the B±W∓ group was removed; however, all 10 *S*s from this latter group are included in the graphical analyses. Similarly, in the between-*S* experiment one *S* from the partial group was discarded due to illness, and the data of the *S* with the same number in the continuous group were removed for the statistical analysis but left in the graphical presentation of the results. Group *N*s were equalized in the same way to allow meaningful comparisons of retrace data between groups.

The results of both experiments were combined to make within- and between-*S* comparisons from the same conditions and together will be referred to as the results of Experiment 4.

Results

The results of Experiment 4 will be presented in somewhat more detail than the others because it allows for the largest number of comparisons. We will begin with a detailed graphical description (Fig. 9) of the within-*S* and between-*S* data both for acquisition and extinction showing the various relationships for all five measures. The summary of an analysis of variance of the data from the within-*S* portion for all measures is reported in Table 1, and from the between-*S* portion in Table 2. Then, in Figure 10, we will present the data from the within-*S* portion of the experiment in a manner which stresses the changing relationships between response strength to S_1 and S_2 that occur across successive segments or measures of the runway. In Figure 11 extinction data are presented for individual *S*s in the within-*S* condition for the goal measure only, and finally, Figures 12, 13, and 14 present a variety of comparisons based on a retrace measure taken for the first time in the between-*S* groups in this experiment. This measure is simply a count of the number of trials on which *S*s retrace in the alley, that is, the number of trials on which they stopped, turned, and went in the opposite direction from the goal.

Figure 9 is a summary of all of the speed data in this experiment. It contains 10 separate acquisition-extinction panels, one for each of five measures for each of the two relationships between color and reinforcement. The acquisition data are plotted in 4-day blocks of trials and the extinction data in 2-day blocks. While there are differences from panel to panel, these data have certain outstanding features. The most prominent single feature is the difference, in all measures, between the continuous between-*S* curve and all of the others, both in acquisition and extinction. In acquisition the indication is that, as in Experiment 3, the between-*S* continuous group shows greater vigor than any of the others, while in extinction the indication is that this group extinguishes more rapidly than any of the others. The other features of note concern within-*S* and between-*S* comparisons, separately.

Within-S acquisition and extinction. The pattern of results in the within-*S* curves, both across color conditions and across performance measures (within color conditions) is very similar to that of Figure 2 of Experiment 1. While the differences are not so pronounced as in the earlier experiment, the indications are the same: the paradoxical effect, faster running to the partial than to the continuous stimulus, is clearer when $S_1\pm$ is white than when it is black, and this difference under the W±B∓ condition disappears and even reverses slightly in the goal measure. When $S_1\pm$ is black (B±W∓) there is little if any indication of faster running to the partial stimulus, but as we go from the start measure to the goal measure there is a gradual divergence of the curves toward faster running to $S_2\mp$. These relationships were made the subject of analyses of variance, a summary of which we present in Table 1. In these analyses, as can be seen from the table, Color Group is a between-*S* factor and Day and Reinforcement are within-*S* factors in the acquisition and extinction analyses. Analyses were performed for each measure at three separate stages of the experiment: from the beginning to the middle of the acquisition period (without the first 2 days), from the middle to

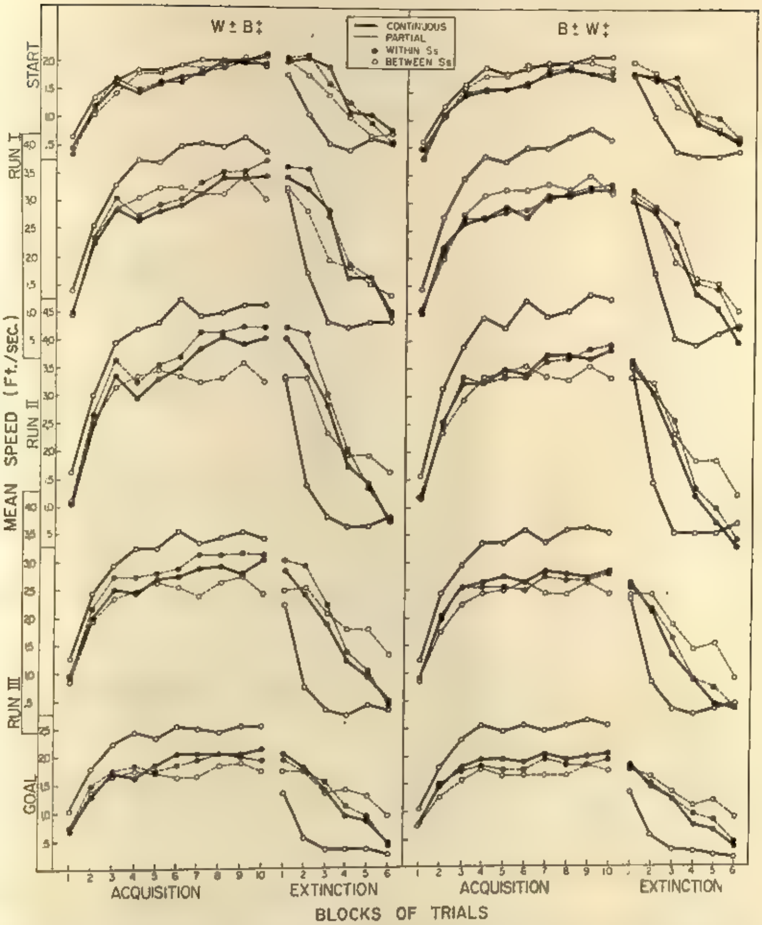

FIG. 9. Within- and between-*S* comparisons of acquisition and extinction performance shown separately for the two color conditions of Experiment 4. Acquisition data are plotted in 4-day blocks of trials (means of eight partial and eight continuous trials) and extinction data are plotted in 2-day blocks.

the end of the acquisition period, and for all of extinction.

In the analyses there was no main effect of Color Group in any measure in any segment of the experiment. There was a reliable effect of Day throughout the experiment, reflecting only the changes which occur in acquisition and extinction. The Color Group × Day effect is significant only in the start measure during the latter half of acquisition.

The pattern of significance related to the Reinforcement factor is one which reflects the graphical differences between groups shown in Figure 9. Reinforcement is significant as either a main effect or interaction in all but the start measure in the first half of acquisition and in the Run III and goal measures during the remainder of acquisition. The Reinforcement main effect fails to reach significance in the Runs I and II measures. The interactions reflect the changes in relationship between the partial and continuous curves that are evident between the groups, especially in the three run measures, there being a clear superiority of the partial curve in the W±B∓ group, with the difference between the curves tending in the opposite direction in the B±W∓ group.

In extinction the triple interaction of Color Group, Day and Reinforcement is significant in the start and Run I measures. These interactions reflect the tendency of

TABLE 1

Summary of Analyses of Variance for Within-S Groups of Experiment 4

Source of variation	Acquisition											
	Days 03–20						Days 21–40					
	df	Start	Run I	Run II	Run III	Goal	*df*	Start	Run I	Run II	Run III	Goal
Color group	1	<1	<1	<1	<1	<1	1	<1	<1	<1	<1	<1
Error 1	16	(7.34)	(24.43)	(37.98)	(17.47)	(8.00)	16	(12.96)	(70.52)	(145.41)	(56.32)	(14.43)
Day	17	35.18***	22.42***	25.84***	26.45***	15.14***	19	6.69***	8.50***	5.96***	4.01***	2.90**
Color Group × Day	17	<1	<1	<1	<1	<1	19	2.95**	1.56	<1	<1	1.44
Error 2	272	(.25)	(.82)	(1.23)	(.65)	(.38)	304	(.14)	(.36)	(.54)	(.23)	(.13)
Reinforcement	1	<1	6.52*	6.94*	4.04	<1	1	<1	4.14	4.10	4.33	7.85*
Color Group × Reinforcement	1	<1	3.29	4.83*	12.73**	7.61*	1	<1	1.16	1.86	13.85**	<1
Error 3	16	(.13)	(.24)	(.63)	(.46)	(.24)	16	(.91)	(1.01)	(1.46)	(.50)	(.42)
Day × Reinforcement	17	<1	<1	1.21	1.00	1.63	19	2.22*	1.13	1.50	<1	1.07
Color Group × Day × Reinforcement	17	<1	<1	<1	1.18	1.59	19	1.33	1.04	<1	<1	1.29
Error 4	272	(.11)	(.23)	(.33)	(.22)	(.13)	304	(.09)	(.23)	(.32)	(.23)	(.14)
Total	647						719					

		Extinction				
Source of variation	*df*	Start	Run I	Run II	Run III	Goal
Color group	1	<1	<1	<1	1.17	<1
Error 1	16	(9.27)	(34.79)	(55.61)	(17.75)	(4.50)
Day	11	60.03***	63.16***	61.92***	60.45***	67.33***
Color Group × Day	11	<1	<1	<1	<1	<1
Error 2	176	(.36)	(1.18)	(2.04)	(1.03)	(.32)
Reinforcement	1	<1	12.45**	15.58**	9.02**	4.94*
Color Group × Reinforcement	1	3.70	2.91	<1	1.29	<1
Error 3	16	(.41)	(.64)	(.60)	(.52)	(.21)
Day × Reinforcement	11	1.20	<1	<1	2.02*	2.15*
Color Group × Day × Reinforcement	11	4.50***	3.28**	1.64	1.52	1.69
Error 4	176	(.14)	(.32)	(.55)	(.33)	(.13)
Total	431					

* $p < .05$.
** $p < .01$.
*** $p < .001$.

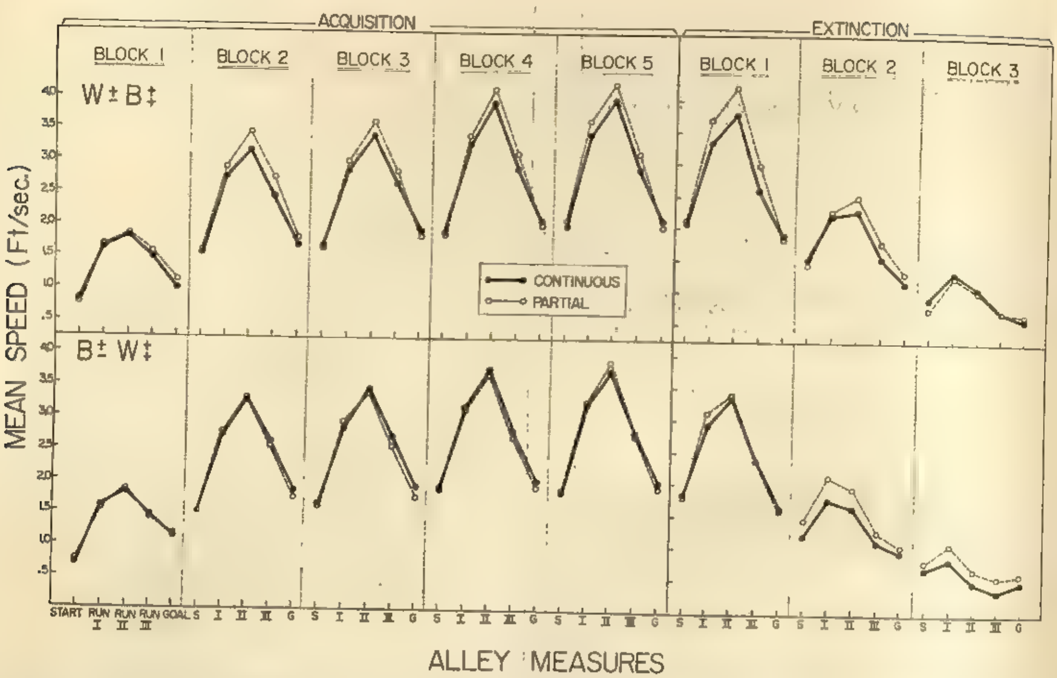

Fig. 10. Within-S acquisition data for the two color conditions of Experiment 4 replotted to emphasize vigor relationships between $S_1\pm$ and $S_2\mp$ in the five segments of the alley. Acquisition data are plotted in 8-day blocks of trials, extinction data in 4-day blocks.

the partial curve to come close to the continuous curve as extinction progresses and even to cross it in the W±B∓ group while the partial curve generally remains above the continuous curve, although parallel to it, in the B±W∓ group. The Day × Reinforcement interaction in the extinction analyses is significant at the .05 level in both the Run III and goal measures, the continuous curve in the goal measure crossing over the partial curve, although the differences are by no means great, especially in comparison to the between-S differences. There is, then, to this mild extent a suggestion of slope differences and, therefore, of the presence of a PRE-like effect. The significant main effect of Reinforcement in extinction seems to reflect only acquisition levels of performance, particularly in the W±B∓ group.

The large color effect observed in extinction of the within-S groups of Experiment 3 was not found here, although there are some indications of color effects in the early measures. However, extinction in Experiment 3 was between-S while extinction in the present experiment was within-S. It is possible that, in Experiment 3, color effects are enhanced because, after acquisition to both colors, extinction trials are to one color only. No significant Color Group × Day interactions were obtained in extinction in Experiment 4 although they were found in Experiments 1 and 2.

To summarize, the analyses of variance suggest that the within-S differences which appear in Figure 9 are reliable; that there are significant partial reinforcement acquisition effects within-S similar to those shown in Experiment 1, and that these seem, again, to depend somewhat on color; and that there are significant differences in extinction due to the reinforcement variable in acquisition, but that these cannot be thought of as depending on anything but the terminal level of the partial and continuous curves in acquisition. The failure to find a consistent or strong Day × Reinforcement interaction reflects little, if any, PRE in the sense of differences in slope. There is, however, the suggestion of such an effect.

Figure 10 is a graphic reanalysis of the within-S data of Figure 9. This entire portion

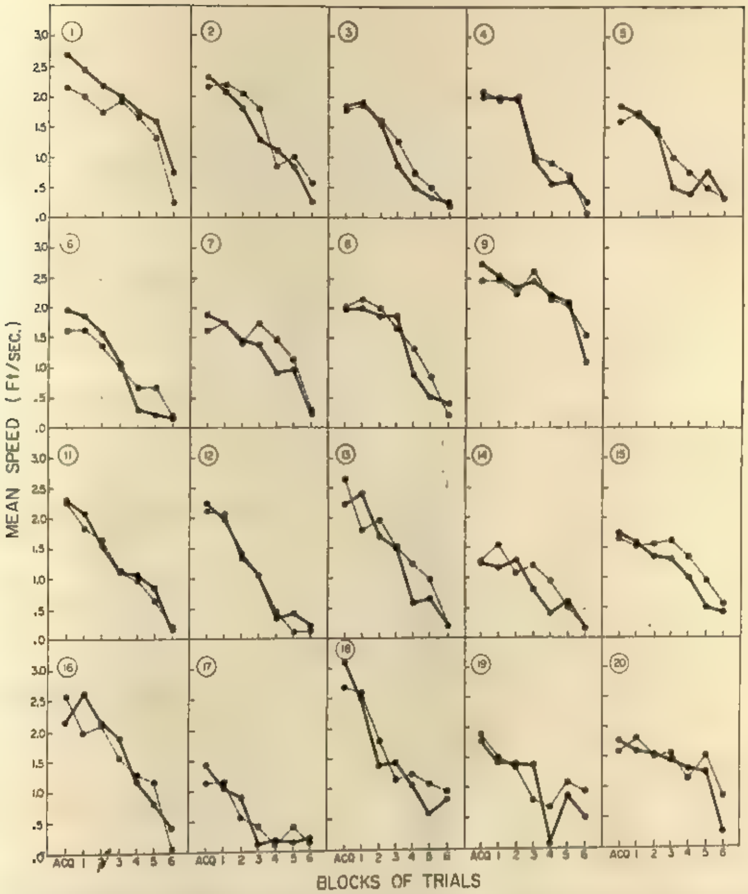

FIG. 11. Individual animal extinction data for the goal measure for *S*s from the within-*S* groups of Experiment 4. The data are plotted in 2-day blocks of trials. The solid curve represents $S_2\mp$ performance, the dotted curve $S_1\pm$ performance.

of the experiment is broken down into eight blocks of trials, five acquisition blocks (each consisting of 8 days) and three extinction blocks (each consisting of 4 days). In each panel, speed of response to $S_1\pm$ and to $S_2\mp$ are plotted against the five successive measures from start to goal. The very subtle changes in the relationship between the strength of response to S_1 and to S_2 over the course of the experiment, and the interaction of these changes in relation to the measure of performance taken, can be seen in a graphic analysis such as this. In such an analysis the characteristics of the data which have been confirmed by the analysis of variance become clearer, particularly in the upper panel describing the condition in which $S_1\pm$ is white and $S_2\mp$ is black. The first block of trials shows the partial stimulus evoking faster running in the goal measure and the adjacent run measure. By Block 2 there is faster running in all measures except for the start measure to $S_1\pm$; by Block 3 the relationship in the goal measures is reversed and $S_2\mp$ now produces higher performance, and these effects hold through Blocks 4 and 5 of acquisition. The general contours of the curves from panel to panel indicate (*a*) that the inverted U-shaped function relating speed to successive measures increases in overall height as we go from Block 1 to Block 5 of acquisition, and (*b*) that the function becomes more peaked as acquisition training progresses, the fastest performance always being in the middle segment of the alley. The curves are somewhat asymmetri-

TABLE 2

SUMMARY OF ANALYSES OF VARIANCE FOR BETWEEN-S GROUPS OF EXPERIMENT 4

Source of variation	Acquisition											
	Days 01–20						Days 21–40					
	df	Start	Run I	Run II	Run III	Goal	*df*	Start	Run I	Run II	Run III	Goal
Reinforcement	1	1.70	12.63**	15.63**	25.13***	56.55***	1	<1	13.80**	18.30***	24.18***	46.96***
Error 1	16	(4.17)	(10.00)	(15.65)	(7.10)	(2.51)	16	(3.50)	(20.91)	(36.58)	(15.82)	(5.07)
Day	19	82.98***	94.13***	83.00***	86.63***	84.76***	19	1.90*	2.92**	2.37**	3.34***	4.53***
Reinforcement × Day	19	1.26	<1	1.47	1.75	3.71***	19	<1	<1	<1	<1	1.16
Error 2	304	(.26)	(.69)	(.96)	(.53)	(.23)	304	(.18)	(.56)	(.70)	(.36)	(.17)
Color	1	1.13	2.16	1.97	2.17	8.95**	1	<1	<1	<1	3.18	1.97
Reinforcement × Color	1	<1	<1	<1	<1	<1	1	5.71*	7.26*	1.65	<1	<1
Error 3	16	(.14)	(.73)	(.48)	(.27)	(.18)	16	(.07)	(.58)	(.91)	(.50)	(.24)
Day × Color	19	1.50	1.82*	1.20	<1	<1	19	2.11**	1.43	2.42**	3.00**	1.93*
Reinforcement × Day × Color	19	1.08	<1	<1	<1	1.00	19	<1	<1	1.04	<1	<1
Error 4	304	(.12)	(.28)	(.40)	(.26)	(.15)	304	(.09)	(.30)	(.26)	(.18)	(.15)
Total	719						719					

Source of variation	df	Extinction				
		Start	Run I	Run II	Run III	Goal
Reinforcement	1	8.54**	11.95**	20.05***	41.73***	82.69***
Error 1	16	(6.21)	(12.30)	(13.42)	(6.62)	(2.01)
Day	11	40.97***	65.06***	46.32***	35.25***	40.86***
Reinforcement × Day	11	3.67***	6.60***	8.91***	9.44***	7.01***
Error 2	176	(.52)	(.87)	(1.25)	(.74)	(.23)
Color	1	1.84	<1	<1	12.25**	<1
Reinforcement × Color	1	1.69	4.66*	<1	6.48*	8.83**
Error 3	16	(.30)	(.52)	(1.02)	(.37)	(.20)
Day × Color	11	1.26	1.15	1.03	<1	<1
Reinforcement × Day × Color	11	<1	<1	1.33	<1	1.21
Error 4	176	(.19)	(.47)	(.66)	(.33)	(.14)
Total	431					

* $p < .05$.
** $p < .01$.
*** $p < .001$.

cal: there is a consistent difference in speed between the first and third run measures, the first showing faster speed than the third; and there are also slightly higher speeds in the goal than in the start measure. These characteristics of the data are similar to those reported by Weiss (1960) in a runway experiment. The picture in extinction is (*a*) of course a decline in the levels of the curves from the first to the third extinction blocks, (*b*) a decline also in the peakedness and symmetry of the functions, and (*c*) a somewhat clearer indication than in Figure 9 of faster running in extinction to the partial stimulus than to the continuous stimulus.

The panels of Figure 11 show individual *S* data for the last block of acquisition trials and for each of the extinction blocks for the goal measure. There is a slight overall tendency for the partial curves (dashed line) to lie somewhat above the continuous ones. The differences are not great, and there are few indications of difference in slope.

*Between-*S *acquisition and extinction.* The differences in the between-*S* curves (Figure 9) are very obvious and clearly replicate the results of Experiment 3 in every respect. First of all, every measure but the start measure shows that the continuous group performs more vigorously than the partial group and that there is no suggestion of the Goodrich type of crossover effect in these between-group data. There is a very clear PRE, the continuous group curve dropping very abruptly at the first and second extinction points, the partial curve showing practically no drop at all at these two points. What continues also to be remarkable in this experiment, as it was in Experiment 3, is the similarity of levels and slopes of the two within-*S* extinction curves and the extinction curve from the partial between-*S* group. There is some suggestion (as there is also in the extinction comparisons of Figure 8) that the between-*S* partial curve reflects greater resistance to extinction than the two within-*S* curves. This shows up particularly in the Run I, Run II, and goal measures in the later extinction trials, where the slope of the within-*S* curves lies somewhat intermediate between the slopes of the between-*S* partial and continuous curves.

Table 2 provides a summary of the between-*S* analyses of variance. Analyses were conducted over Days 1–20 (Blocks 1–10 in Figure 9), Days 21–40 (Blocks 11–20 in Figure 9), and over the extinction days for each measure. In these analyses Reinforcement is a between-*S* factor and Day and Color are within-*S* factors. The table shows that Reinforcement is significant in both segments of acquisition in all but the start measure where it does not reach significance in either segment. This effect reflects the obvious acquisition differences shown in Figure 9 between partial and continuous performance. There is a significant Day effect in all measures in both segments of acquisition, and a Reinforcement × Day effect in the first segment of acquisition in the goal measure. The Day effect reflects continuing changes throughout acquisition and the interaction reflects a slower rate of acquisition of the response by the partial group in the goal measure.

In the first segment of acquisition a main effect of Color appears only in the goal measure, and there is a Day × Color interaction at the .05 level in the first running measure. Several color interactions occur in the second segment of acquisition: Reinforcement × Color in the start and Run I measures, and Day × Color in all but the Run III measure. These color effects do not show any particularly consistent pattern either across measures or across acquisition segments, but Figure 9 does reflect a tendency for performance to the white stimulus to be more vigorous when it is partially rewarded and to be less vigorous when continuously rewarded, especially in the second segment.

In extinction the Reinforcement, Day, and Day × Reinforcement effects are all significant in every measure, the latter effect reflecting differential slopes of partial and continuous curves and thereby substantiating what is already obvious from Figure 9—that a PRE occurred in the between-*S* groups. Color appears as a significant effect in extinction in a color main effect in Run III and in Color × Reinforcement interactions in the Run I, Run III and goal measures. These color effects reflect somewhat more rapid extinction of the CRF group in

the presence of the white stimulus. These same color effects are apparent in the retrace data.

In general, the analyses of variance support the pattern of relationships shown in Figure 9.

Retrace data. In this experiment retrace data were taken only for the between-*S* groups. These data are presented here because they suggest a method of measuring extinction performance we had not used before, a method which provided such a clear-cut indication of the differences between the CRF and PRF groups in extinction that we decided to use it in later experiments for within- as well as between-*S* analyses. Figure 12 shows number of retraces by the continuous (C) and partial (P) groups in 2-day blocks over the entire extinction phase of the experiment. (No retracing was observed in either group except on the first few trials during acquisition.) In this figure the maximum number of retraces at each block is 72 which would be reached only if all *S*s retraced on every trial. It is very evident that the CRF group retraces on far more trials than does the PRF group during extinction, and that frequency of retracing in the continuous group appears to increase in a somewhat negatively accelerated manner starting with the first 2-day block of trials. No retracing was observed in the partial group until the fourth block of trials. At the point at which we terminated extinction, the partial group was retracing at about the level of the continuous group at the first block.

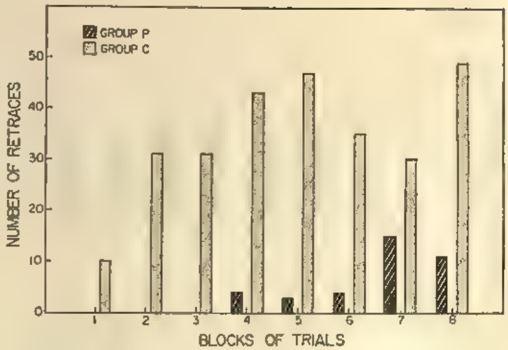

FIG. 12. Number of retraces in extinction by between-*S* groups of Experiment 4. Data are based on 2-day blocks of trials. Only the first retrace on any trial was scored for each animal and a maximum score (all *S*s retracing on all trials) for a 2-day block is 72.

Figure 13 shows the data for retracing broken down according to the five measures taken in the runway. The most noteworthy feature of these data is that the retracing occurs primarily in the first two segments, with less in the last two segments and very little in the middle one. Early in extinction retracing is seen first in the goal region and continues to occur there; but as training progresses, occurrence of retracing tends to be more and more in the first segments of the runway. Late in extinction there is less retracing at the start and increased retracing in the goal segment.

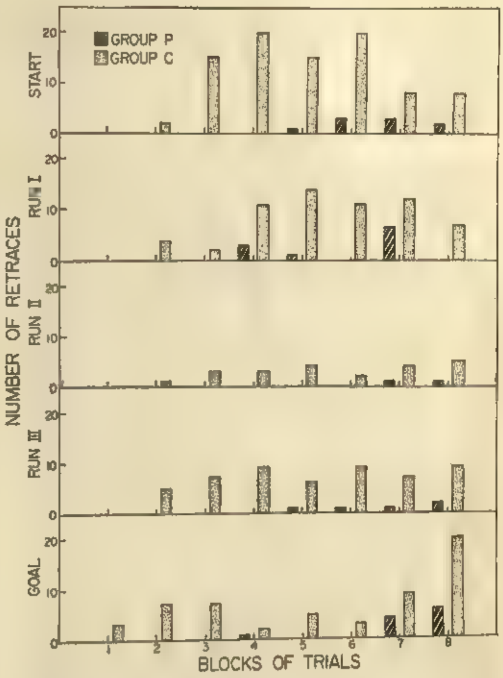

FIG. 13. Number of retraces by between-*S* groups of Experiment 4 replotted to show segment of alley in which retraces occurred.

Finally, Figure 14 breaks retracing down according to the color of the alley in which it occurs. There is some suggestion that the *S*s tend to retrace more in the white alley than they do in the black. At least this is a consistent picture occurring in every one of the 2-day blocks for the CRF *S*s. However,

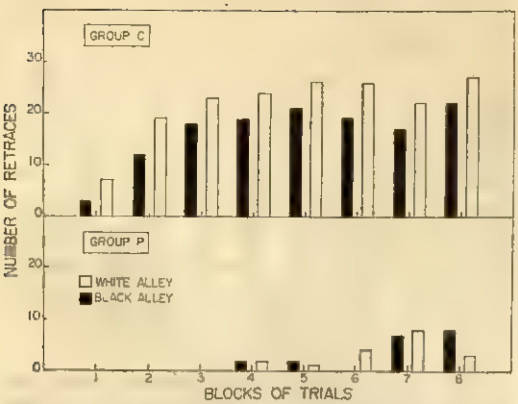

FIG. 14. Number of retraces by between-S groups of Experiment 4 replotted to show black and white alley retraces.

for each of the alley colors, separately, there is an increase of about the same form in retracing from the beginning to the end of extinction.

DISCUSSION

The experiments were performed to extend the generality of a finding (Amsel et al., 1964) that PRF acquisition effects are discernible in within-S experiments; to investigate extinction effects following within-S acquisition; and to compare results obtained under within-S and between-S conditions in both acquisition and extinction. In making these comparisons, we came across some interesting and unexpected findings, and these have raised a number of questions which can only be answered by further work, some of which is now under way in our laboratory.

Absence of a "PRE" within-Ss. Why is there little, if any, evidence for a difference in extinction to two stimuli, one of which has been related to PRF and the other to CRF in within-S partial reinforcement acquisition? In terms of our within- versus between-S comparisons, this question becomes: Why does extinction to each of the two stimuli look more like extinction after partial than like extinction after continuous reinforcement? Two answers to this question can perhaps be provided by the theory with which we have been working (Amsel 1958, 1962), and one of these is shown schematically in Figure 15. The argument diagrammed in the figure is that the within-S case permits generalization of the persistence effect (r_F—s_F → approach) from S_1 to S_2 in extinction; that is, as soon as either stimulus elicits r_F in extinction, the s_F stimulus cues in the persistence mechanism regardless of whether the external stimulus has been associated with PRF or CRF training in acquisition. Since the within-S case permits mediated generalization of the mechanism for the PRE, we might expect PRE-like effects to both stimuli in extinction. Of course, the between-S case does not permit such generalization—PRF and CRF acquisition develop in separate organisms, and transfer of persistence effects is therefore impossible under between-S conditions. We are suggesting that the critical factor determining resistance to extinction is whether s_F elicits approach (as well as avoidance). It would seem, if our reasoning is correct, that this internal control of extinction behavior was the important factor in our within-S experiments, and that differential external stimulation was relatively unimportant, even though other aspects of the data make it clear that Ss responded differentially to S_1 and S_2 in acquisition.

Spear (1964) and Spear and Pavlik (1966) have reported extinction speed data from choice experiments which are consistent with our data and with an explanation in terms of mediated generalization. In these experiments, Ss were given equal experience with partially and continuously rewarded arms of a T-maze and were then extinguished in both arms of the maze. The extinction data showed PRE-like speeds in the continuously rewarded arm of the maze, and Spear (1964) suggests that the mechanism producing the PRE in the continuous arm appears to be independent of the external stimulus situation. These findings are similar to those reported in the present experiments, and the theoretical schema provided in Figure 15 makes explicit a mechanism for producing these effects which is independent of external stimulus factors.

Another answer to the question of why extinction performance is the same to $S_1\pm$ and S_2+ involves the concept of primary

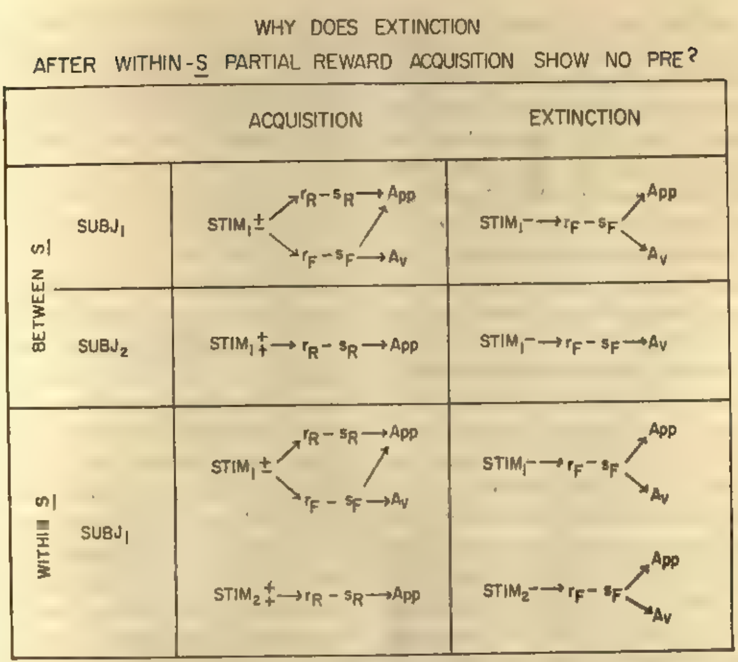

FIG. 15. A comparison of factors operating to affect extinction following between-*S* and within-*S* acquisition.

stimulus generalization of $r_F - s_F$ in acquisition. This second possibility, detailed elsewhere (Amsel, 1966) holds that, if S_1 and S_2 are sufficiently similar that r_F occurs not only to $S_1\pm$ during acquisition but generalizes to $S_2\mp$, then s_F will become conditioned to the S_2 response. We would then have no basis for expecting a difference in rate of extinction to S_1 and S_2. While this explanation clearly differs from the mediated (or secondary) generalization explanation outlined in Figure 15, the two are by no means incompatible; both mechanisms might operate to produce the within-*S* extinction result.

The recent data of Brown and Logan (1965) are similar in their import to the extinction data reported in this monograph. Their explanation for what they call the "generalized partial reinforcement effect," following Logan, Beier, and Kincaid (1956), involves primary generalization of learned responses to "stop $r_G$ing" from the partial to the continuous stimulus as the mechanism which accounts for the within-*S* extinction result. This explanation differs from our primary generalization explanation in that (*a*) the "stop $r_G$ing" response is not anticipatory in character but, rather, occurs *after* the presentation of the goal event, and (*b*) this generalization occurs in extinction and not during acquisition. While we have also proposed a possible extinction interpretation of the generalized PRE, it involves mediated generalization rather than the primary generalization of the Brown and Logan explanation.

We have embarked on some follow-up experiments to provide more information about the within-*S* extinction effect and its explanation. One approach we have tried is to vary, not only a single stimulus in an otherwise constant situation, but to vary grosser aspects of the stimulus situation even to the extent that approach responses to the differing situations will take different forms.

An experiment on within-*S* PRF training and extinction (Rashotte, 1966) employed pairs of responses which differed in similarity in an attempt to increase the within-*S* PRE. There were two conditions in the experiment. In the first condition (R_1/R_2) rats were trained to make similar running responses to approach food in two apparatuses which differed not only in the

black-white dimension but also in other respects (e.g., width of runway). In the second condition (C/R_2) *S*s were trained to perform two different responses to approach food: climbing in a black apparatus and running in a white alley. In both cases one S-R sequence was continuously rewarded, the other partially, and there were groups within each condition counterbalanced for the response-reward contingency. Between-*S* 'control' groups were also included in the experiment. The extinction data from this experiment show an unmistakable within-*S* PRE, the difference between the conditions being that the PRE develops later in R_1/R_2 than in C/R_2. The between-*S* controls (each group also makes two responses, e.g., $C\mp R_2\mp$ versus $C\pm R_2\pm$) also show the usual PRE, of greater magnitude than that found in within-*S* extinction after within-*S* acquisition. While this experiment does not allow us to choose between the primary and mediated generalization explanations, it does indicate that PRF-like extinction is not a necessary characteristic of responses acquired under continuous reward conditions in the within-*S* experiment.

Our other approaches to investigating the within-*S* extinction phenomenon have involved the same apparatus as employed in the four experiments presented in this report. One idea has been to reduce generalization by separating out, into separate trial-time blocks, the black and white stimuli: only the black alley trials are given at one time of day and only white alley trials at another; one type of trial is always conducted by one experimenter, under specific background-noise conditions, while the other type of trial is given by a different experimenter in the presence of a different background noise. We are currently conducting this experiment under the conditions that, for example, all of the $S_1\pm$ trials (e.g., PRF trials to the black stimulus) are run in the morning, and all of the $S_2\mp$ trials (CRF trials to the white stimulus) are run 12 hours later, in the evening. A procedure such as this is very similar to some of the variety of classical conditioning experiments on "switching" reported by Asratian (1965) which appear to be successful in reducing stimulus generalization. Some of these involve establishing a conditioned response at two separate times, such as morning-afternoon, to the same conditioned stimuli but different unconditioned stimuli. In one experiment, to quote Asratian, they demonstrated "that it is possible to form a positive and a negative conditioned reflex of the same kind to a single stimulus, i.e., that it is possible to switch conditioned reflexes from one functional sign to the opposite sign within the limits of a single type of activity of the organism. . . ." The same investigator conducted two experiments with the same dogs in the same chamber, one in the morning and one in the afternoon, and the difference was only that in the morning all of the stimuli were reinforced, while in the afternoon one of the stimuli was not reinforced. They found that this stimulus acted as a positive CS in the morning, but like a negative CS in the afternoon. These same kinds of experiments were also performed with two different strengths or delays of UCR, each connected to a different conditioning chamber, to a different *E*, or to different times of day, e.g., morning-afternoon. The problem in these Pavlovian experiments is whether different strengths of CR can thus be conditioned to "the same CS," that is to say, whether the dog can learn different intensities or delays of the same conditioned response to the same conditioned stimulus when other "background" stimuli (the chamber) and/or the time of day are varied from one session to another.

A third procedure for investigating the source of our within-*S* extinction finding involves a technique of predifferentiation of S_1 and S_2. We are conducting an experiment in which *S*s are first trained in a black-white discrimination in our apparatus on the basis of nonreinforcement to one stimulus (S_1-) and reinforcement to the other (S_2+). After this differentiation has been learned, the stimulus associated with nonreward will be switched to partial reward and a prolonged period of $S_1\pm S_2\mp$ training ensues (a comparison group will be switched to $S_1\mp S_2\pm$) followed, in turn, by within-*S* extinction. We wonder whether the predifferentiation procedure will separate the action of the

exteroceptive stimulation sufficiently to enable these stimuli to exercise greater differential control over extinction behavior and produce the PRE. Such a procedure has, in effect, been employed by Stein (1957) in an experiment which might be regarded as demonstrating the within-*S* PRE in relation to the conditioned emotional response under free-responding conditions.

A fourth approach to a better understanding of the within-*S* PRE is perhaps the simplest and most straightforward: variation in percentage reward to $S_1\pm$ in within-*S* PRF acquisition prior to within-*S* extinction. True, a clear within-*S* PRE has been demonstrable in the Rashotte experiment when different responses, as well as different stimuli, were employed on the $S_1 50\%$–$S_2 100\%$ schedule. Nevertheless, failure to demonstrate the effect convincingly when only one dimension of stimulation (black-white) serves as the differential cue may reflect the particular reinforcement percentage of S_1 employed. In a recent study, Henderson (1966) gave within-*S* acquisition training to groups run under a variety of S_1 percentage reinforcement conditions (0, 12, 25, 50, 100) holding S_2 constant across groups at 100%. The finding was that response speeds to $S_2\mp$ varied directly with percentage reward to $S_1\pm$. Unfortunately for our present purposes Henderson did not run within-*S* extinction in the second phase of the experiment, but instead switched to a variant of discrimination reversal (S_1+S_2-). We are presently in the process of conducting the within-*S*-extinction version of this experiment. The Henderson acquisition finding does, incidentally, suggest that there is a generalization of inhibitory factors from $S_1\pm$ to $S_2\mp$ during acquisition.

An experiment by Davenport (1963b) in which *S*s were trained to run to two stimuli, one associated with 100% reward for all groups, the other with either 67, 33 or 0%, reports free choice data and also speeds obtained on forced trials to the two stimuli. These speed data of Davenport's also suggest generalization of inhibitory factors from the lesser percentage stimulus to the 100% stimulus. These findings of Henderson and Davenport lend support to the primary-generalization-of-r_F explanation proposed earlier.

Pavlik and his co-workers have reported a variety of within-*S* extinction findings both from free-operant and from runway situations. The data from the free-operant experiments have reflected both a reversed PRE (Pavlik & Carlton, 1965), and a conventional PRE (Pavlik, Carlton, & Manto, 1965). The former result was obtained when one lever was employed and time of exposure to S_1 and S_2 (red and white arc lights) was equalized; the latter result was obtained both when number of responses and number of reinforcements to S_1 and S_2 (rather than time of exposure) were equalized. In the runway experiment (Pavlik, Carlton, & Hughes, 1965), a *reversed* PRE was obtained only in a goal measure which extended over the last 4 ft. of a 6 ft. runway, rate of extinction to the partial and continuous stimuli being the same in the start measure. As we have mentioned earlier the free-operant and discrete-trial experiments require different explanatory emphases, and we will not attempt to pursue these here. The difference between runway results of Pavlik, Carlton, and Hughes and our own (as well as those of Brown & Logan) may well be due to procedural differences noted in their discussion, particularly intertrial interval, relatively long in our experiments and very short in theirs.

Greater vigor to $S_1\pm$ early in training near the goal. A second way in which our within-*S* results look different from our between-*S* results is in relation to performance on early acquisition trials. All of the earlier between-*S* results and our own current ones are in agreement that, in the goal segment, *S*s under CRF conditions show greater vigor of performance than *S*s under PRF conditions from the very outset of training, and that this difference develops and stabilizes without reversal. Some of our within-*S* data suggest that $S_1\pm$ elicits greater response vigor in the goal region early in training than does $S_2\mp$ (or S_2+), although in most of our experiments the relative terminal levels of performance to S_1 and S_2 in the goal region correspond to the between-*S* case, the partial stimulus eliciting lesser vigor than the con-

tinuous stimulus at that stage. It would be unwise to overplay the importance of this suggestion from our data; the effect is not statistically reliable in Experiments 1, 2, or 4 and in Experiment 3 it appears to depend on Color, as witness, for example, the Color Group × Reinforcement interaction in the statistical analysis of the goal data for the early part of acquisition. However, this same kind of effect seems to occur in several other experiments in our laboratory (e.g., Henderson, 1966), and the weight of its appearance in a number of studies points to its possible reliability. Should the early facilitation of responses to $S_1\pm$ prove to be a reliable phenomenon, the mechanism of anticipatory frustration might explain both this effect and the more vigorous performance of partially reinforced responses observed early in the response chain late in acquisition training. Such an explanation might go something like this: in the within-*S* experiment, r_R building up to $S_2\mp$ generalizes strongly to $S_1\pm$ and creates, much earlier than is possible in a between-*S* PRF group, the conditions which are necessary for nonreward to produce R_F. At this early stage, then, $S_1\pm$ and $S_2\mp$ evoke reactions which are different in this sense: that, to $S_1\pm$ but not to $S_2\mp$, as *S* gets closer to the goal, early in training, there is aroused a mild r_F reaction, too weak to be aversive but strong enough to be mildly, and unspecifically, exciting. There would be little reason to expect r_F to generalize to $S_2\mp$ at this stage since (*a*) r_F is weak, and (*b*) gradients of r_F are steeper than are gradients of r_R (Amsel, 1962). As training continues, the strength of r_F increases to the point of aversiveness, and responding becomes less vigorous in the goal region to $S_1\pm$ than to $S_2\mp$. When r_F is strong at the goal and generalizes weakly to the run (and even to start) segments, it serves as a nonspecific energizer to produce the acquisition crossover in the early segments of the response chain. In the within-*S* experiment we might therefore expect the same kind of "crossover" near the goal *early in training* as we see in some of the earlier alley segments *later in training*.

According to this line of reasoning, the degree of initial facilitation of responding to $S_1\pm$ should be a function of the amount of r_R present early in training when nonrewards occur in the presence of S_1. We have tried to manipulate the strength of r_R by direct goal-box placements preceding within-*S* PRF acquisition; but our first attempt has been unsuccessful. A recent experiment by Trapold and Doren (1966) suggests that a replication of our experiment with more careful attention to the details of the placement procedure might be in order. They showed that the exact position of placement of an *S* in a goal box made a crucial difference; that when, on partial-reward placement trials, *S* was placed directly over the food cup no PRE was observed on subsequent extinction running trials, while placement of *S* 8 in. from the food cup requiring a short approach response, yielded a PRE from PR placements.

PRF acquisition effects within-Ss. The within-*S* acquisition data in our experiments are, for the most part, successful replications of our earlier acquisition findings (Amsel et al., 1964), which have also recently been replicated in another laboratory (Ludvigson, 1966). Attention should be drawn to the fact that in Experiments 3 and 4, in which acquisition comparisons can be made between within-*S* and between-*S* curves, it is clear that the within-*S* differences are smaller than those found in the between-*S* condition, and that this discrepancy occurs mainly because the within-*S* continuous curve is lower than its between-*S* counterpart. This finding suggests generalization of inhibitory factors from $S_1\pm$ to $S_2\mp$ during acquisition, a finding also confirmed in a later experiment (Henderson, 1966).

PRF acquisition effects between-Ss. Why is there no obvious Goodrich-type effect in the running or starting data of our between-*S* experiments? It is very surprising that such a question would need to be asked. We certainly did not expect that the type of crossover found by Goodrich and others in the start and run measures of between-*S* PRF experiments would not reappear in our own experiments, and we are faced with an unusual situation. We started out by doing an experiment in which we observed, within-*S*, some of the kinds of phenomena

that Goodrich and others had observed in between-*S* experiments. However, when we ran the between-*S* control which we thought appropriate for our own type of experiment, we could not recover the paradoxical between-*S* asymptotic crossover effect we had taken for granted to begin with. On the other hand, the extinction effect, the PRE, is clear enough in our between-*S* experiments in all measures but is, to say the least, not very clear in the within-*S* experiments.

We can think of two possible variables which may account for our failure to produce crossover effects in between-*S* acquisition. The first and most obvious one is that our between-*S* experiments involved two colors ($S_1\pm S_2\pm$ versus $S_1\mp S_2\mp$) while the earlier ones involved only one ($S_1\pm$ versus $S_1\mp$). It is sheer conjecture to say that the inhibitory effect of partial reward, which always shows up early in the goal area, moves forward in the instrumental sequence more readily when *S* is exposed ("randomly") to two colors of alley rather than just one. This is a possibility without any theoretical connotation of moment, unless running to two different colors in an alley somehow facilitates generalization, particularly of inhibitory effects. The second difference between our procedure and those of Goodrich, Wagner, and others is that while our procedure involves no prior feeding in the goal box and no preliminary adjustment training of any kind, their procedures typically involve a substantial amount of goal box exposure and feeding. The relevance of these variables to the effect in question is being tested in our laboratory, but the results are, as yet, inconclusive.

SUMMARY

In a previous experiment it was reported that when an individual *S* is trained to approach two stimuli, one of which ($S_1\pm$) is associated with partial reward and the other (S_2+) with continuous reward, PRF acquisition effects are obtained which have many similarities to PRF effects obtained in the usual between-*S* experiments. The results of this earlier experiment were that the between-*S* PRF effect, faster running by a partial than a continuous group early in the response chain and slower running by the partial group late in the chain, could be reproduced within the same *S*. This experiment was deficient in several respects: (*a*) black was always the partial stimulus and white the continuous (B±W+) and the reverse color-reinforcement relationships were not included in the design; (*b*) there were no between-*S* groups in the experiment for direct within- and between-*S* comparisons; (*c*) rewards, rather than trials, to each of the stimuli were equated; and (*d*) extinction was not carried out and only PRF acquisition effects were observed. Our present report involves four experiments which cover the deficiencies of the earlier experiment while they allow further analysis of within-*S* and between-*S* PRF experiments.

Experiment 1, which was in part a replication of the earlier experiment, included both color groups as well as extinction. It was found that the PRF acquisition effects reported previously were replicable; however, when *S*s were extinguished to both S_1 and S_2 there was no reliable difference in performance to the stimuli. There was a suggestion that some *S*s extinguished more rapidly to S_2+ than to $S_1\pm$, the usual between-*S* finding. It was further observed that, early in training, *S*s performed more vigorously to $S_1\pm$ than to S_2+ late in the response chain, a result unlike anything found in between-*S* PRF experiments. There was also an indication of a Color × Reinforcement interaction, white interacting with partial reward to produce approach responses of greatest vigor.

In Experiment 2 trials to both stimuli (rather than rewards) were equated ($S_1\pm S_2\mp$), and both acquisition and extinction training were carried out. The results indicated that PRF acquisition and extinction effects under these conditions were similar to those obtained in Experiment 1.

In Experiments 3 and 4 within-*S* groups ($S_1\pm S_2\mp$) as well as between-*S* groups were given acquisition and extinction training with trials to both stimuli equated. The between-*S* groups also ran to both stimuli, but the continuous group was on a CRF schedule to both ($S_1\mp S_2\mp$) while the partial

group was on a PRF schedule to both ($S_1 \pm S_2 \pm$). Besides direct within- and between-*S* comparisons, these experiments allow a comparison of two extinction procedures: in Experiment 3 within-*S* extinction followed both within- and between-*S* acquisition, that is to say, half of the *S*s in each group were extinguished only to $S_1 \pm$ and the other half only to $S_2 \mp$. Comparisons of PRF and CRF extinction performance are here made under between-*S* conditions after both within-*S* and between-*S* acquisition. In Experiment 4 extinction was within-*S* in both conditions, all *S*s being extinguished to both S_1 and S_2.

In both Experiments 3 and 4 the within-*S* groups showed some of the PRF acquisition effects observed in earlier experiments. The between-*S* acquisition data, on the other hand, did not show the often-reported (paradoxical) faster running by the PRF group early in the response chain, but rather showed slower running in PRF acquisition in all segments of the response chain. In Experiment 3 in which extinction was between-*S* there was a strong effect of stimulus intensity in the within-*S* groups such that extinction to white produced a relatively higher level of responding than did extinction to black. Rate of extinction was not affected by stimulus intensity, however, and the within-*S* extinction findings of Experiments 1 and 2 were replicated. In the between-*S*s groups there was also an effect of stimulus intensity, but the usual PRF acquisition effect appeared. The within-*S* extinction performance was much more like between-*S* PRF extinction than like CRF extinction.

In Experiment 4, in which extinction was within-*S*, the effect of stimulus intensity in extinction was reduced and the same pattern of within- and between-*S* results was found as in Experiment 3. A comparison of the extinction performance of within-*S* and between-*S* groups in Experiment 4 shows within-*S* extinction following within-*S* training to lie intermediate between the rates of PRF and CRF between-*S* extinction, being somewhat more like partial than continuous.

The concluding discussion points up the acquisition and extinction characteristics of the within-*S* experiments which differ from those of between-*S* experiments and suggests interpretations of these differences in terms of frustrative nonreward mechanisms. Some ways of investigating further the factors responsible for the within-*S* PRF extinction effect are discussed, and some possible reasons are advanced for failure to obtain the usual between-*S* acquisition effects in these experiments.

REFERENCES

AMSEL, A. The role of frustrative nonreward in noncontinuous reward situations. *Psychological Bulletin*, 1958, 55, 102–119.

AMSEL, A. Frustrative nonreward in partial reinforcement and discrimination learning: Some recent history and a theoretical extension. *Psychological Review*, 1962, 69, 306–328.

AMSEL, A. Partial reinforcement effects on vigor and persistence: Advances in frustration theory derived from a variety of within-subjects experiments. In K. W. Spence, J. T. Spence, & N. H Anderson (Eds.), *The psychology of learning and motivation: Advances in research and theory*. New York: Academic Press (in press).

AMSEL, A., MACKINNON, J. R., RASHOTTE, M. E., & SURRIDGE, C. T. Partial reinforcement (acquisition) effects within subjects. *Journal of the Experimental Analysis of Behavior*, 1964, 7, 135–138.

AMSEL, A., & WARD, J. S. Frustration and persistence: Resistance to discrimination following prior experience with the discriminanda. *Psychological Monographs*, 1965, 79, (4, Whole No. 597).

ASRATIAN, E. A. *Compensatory adaptations, reflex activity, and the brain*. London: Pergamon Press, 1965.

BECK, S. B. Eyelid conditioning as a function of CS intensity, UCS intensity, and manifest anxiety scale score. *Journal of Experimental Psychology*, 1963, 66, 429–438.

BROWN, R. T., & LOGAN, F. A. Generalized partial reinforcement effect. *Journal of Comparative and Physiological Psychology*, 1965, 60, 64–69.

DAVENPORT, J. W. Spatial discrimination and reversal learning involving differential magnitude of reinforcement *Psychological Reports* 1963, 12, 655–665. (a)

Davenport, J. W. Spatial discrimination and reversal learning based upon differential percentage of reinforcement. *Journal of Comparative and Physiological Psychology*, 1963, **56**, 1038–1043. (b)

Ferster, C. B., & Skinner, B. F. *Schedules of reinforcement*. New York: Appleton-Century-Crofts, 1957.

Goodrich, K. P. Performance in different segments of an instrumental response chain as a function of reinforcement schedule. *Journal of Experimental Psychology*, 1959, **57**, 57–63.

Grice, G. R., & Hunter, J. J. Stimulus intensity effects depend upon the type of experimental design. *Psychological Review*, 1964, **71**, 247–256.

Helson, H. *Adaptation level theory*. New York: Harper & Row, 1964.

Henderson, K. Within-subjects partial-reinforcement effects in acquisition and in later discrimination learning. *Journal of Experimental Psychology*, (in press).

Humphreys, L. G. The effect of random alternation of reinforcement on the acquisition and extinction of conditioned eyelid reactions. *Journal of Experimental Psychology*, 1939, **25**, 141–158. (a)

Humphreys, L. G. Acquisition and extinction of verbal expectations in a situation analogous to conditioning. *Journal of Experimental Psychology*, 1939, **25**, 294–301. (b)

Humphreys, L. G. Extinction of conditioned psychogalvanic responses following two conditions of reinforcement. *Journal of Experimental Psychology*, 1940, **27**, 71–75.

Humphreys, L. G. The strength of a Thorndikian response as a function of the number of practice trials. *Journal of Comparative and Physiological Psychology*, 1943, **35**, 101–110.

Jenkins, W. O., & Stanley, J. C. Jr. Partial reinforcement: a review and a critique. *Psychological Bulletin*, 1950, **47**, 193–234.

Logan, F. A., Beier, E. M., & Kincaid, W. D. Extinction following Partial and varied reinforcement. *Journal of Experimental Psychology*, 1956, **52**, 65–70.

Ludvigson, H. W. Differential conditioning and subsequent choice: Partial vs. continuous reward. *Psychonomic Science*, 1966, **4**, 391–392.

Pavlik, W. B., & Carlton, P. L. A reversed partial-reinforcement effect. *Journal of Experimental Psychology*, 1965, **70**, 417–423.

Pavlik, W. B., Carlton, P. L., & Hughes, R. A. Partial reinforcement effects in a runway: Between- and within-*S*s, *Psychonomic Science*, 1965, **3**, 203–204.

Pavlik, W. B., Carlton, P. L., & Manto, P. G. A further study of the partial reinforcement effect within-subjects. *Psychonomic Science*, 1965, **3**, 533–534.

Pavlov, I. P. *Conditioned reflexes*. (Translated by G. V. Anrep) London: Oxford University Press, 1927.

Rashotte, M. E. Frustrative factors in persistence: within- and between-*S* comparisons. Unpublished doctoral thesis, University of Toronto, 1966.

Skinner, B. F. *The behavior of organisms: an experimental analysis*. New York: Appleton-Century, 1938.

Spear, N. E. Choice between magnitude and percentage of reinforcement. *Journal of Experimental Psychology*, 1964, **68**, 44–52.

Spear, N. E., & Pavlik, W. B. Percentage of reinforcement and reward magnitude effects in a T maze: Between and within subjects. *Journal of Experimental Psychology*, 1966, **71**, 521–528.

Trapold, M. A., & Doren, D. G. Effect of noncontingent partial reinforcement on the resistance to extinction of a runway response. *Journal of Experimental Psychology*, 1966, **71**, 429–431.

Spence, K. W. *Behavior theory and learning*. Englewood Cliffs, N. J.: Prentice-Hall, 1960.

Stein, L. The partial reinforcement effect in aversive conditioning. Paper read at Eastern Psychological Association, New York, 1957.

Wagner, A. R. Effects of amount and percentage of reinforcement and number of acquisition trials on conditioning and extinction. *Journal of Experimental Psychology*, 1961, **62**, 234–242.

Weiss, R. F. Deprivation and reward magnitude effects on speed throughout the goal gradient. *Journal of Experimental Psychology*, 1960, **60**, 384–390.

(Received January 14, 1966)

Vol. 80, No. 21 Whole No. 629, 1966

Psychological Monographs: General and Applied

A BLOCK ROTATION TASK:

THE APPLICATION OF MULTIVARIATE AND DECISION THEORY ANALYSIS FOR THE PREDICTION OF ORGANIC BRAIN DISORDER[1]

PAUL SATZ[2]

University of Florida

A multivariate instrument, designed to detect the likelihood of brain disorder, was standardized and repeatedly cross validated. A block rotation test was constructed to measure Ss ability to reproduce block designs as they would look if rotated 90 degrees from the stimulus designs. 6 measures of error were inserted into the discriminant function along with age and WAIS PIQ. The final restandardization was based on 157 brain-injured Ss and 210 controls. Predictive validity was high on each of the validation and cross-validation samples. Discriminant scores were related to type and classification of brain injury, but not to area. The effects of base rates and cost efficiency on decisions were examined.

In the last decade, increasing demands have been made on clinical psychologists to determine the presence or absence of brain lesions in man. Although the success of this venture has been disappointing, the psychologist has continued to predict the likelihood of brain dysfunction from psychological tests. Several reasons, both methodological and theoretical, have been advanced to explain these shortcomings. Methodological criticism (Satz, 1963; Yates, 1954a) has focused on (*a*) failure to employ objective scoring procedures and quantifiable analysis of data; (*b*) failure to use adequate control groups; (*c*) reluctance to report normative data and optimal cut-off points for classification; (*d*) failure to control for the relevant variables of age and intelligence; (*e*) tendencies to report the discriminatory efficiency of a test in terms of group differences rather than classification accuracy; (*f*) failure to employ multivariate statistical procedures or cost efficiency methods to determine the utility of the test(s) in different base rate populations; and (*g*) the lack of adequate cross-validational studies.

Additional criticism has been directed at the failure to consider possible dimensions of the concept "brain disorder." It has been shown that the concept of brain disorder, as a homogeneous diagnostic construct, is at variance with current teachings in neurology (Merritt, 1959) and the experimental findings in psychology (Meyer, 1961; Reitan, 1962). Nevertheless, psychologists have often attempted to construct tests predictive of "brain damage" without examining the effects of possible dimensions within the concept. A few examples are

[1] The first part of this study was presented at the annual meeting of the American Psychological Association, Philadelphia, 1963, and based upon a doctoral dissertation submitted to the Department of Psychology, University of Kentucky, in partial fulfillment of the requirements for the PhD degree. The dissertation was selected as one of the winners in the 1962–1963 Creative Talent Awards, American Institutes of Research, Washington, D.C. The second part of the study (cross validation) was conducted while on a National Institutes of Mental Health Postdoctoral Fellowship and later as an assistant professor at the University of Florida.

[2] The author wishes to express his gratitude to the members of his dissertation committee: Jesse G. Harris, Jr. (Chairman), Frank A. Pattie, Graham B. Dimmick, Frank Essene, and Charles F. Diehl. Particular gratitude is extended to the chairman, Jesse Harris, for his encouragement and stimulating criticism throughout the project, and for his desire to see the cross-validation studies carried out (1963–1966). The author is also deeply grateful to James Calvin for his many stimulating lectures on multivariate statistical models and decision theory.

types of brain disorder (acute, chronic, static), *classification* of lesions (e.g., vascular, neoplastic, traumatic), and *localization* or *lateralization* of lesions (e.g., left or right hemisphere).

Types and Classification of Brain Disorder

Neurological evidence indicates that different types of brain lesions may cause marked differences in both symptomatology and behavior (Merritt, 1959). According to Fitzhugh, Fitzhugh, and Reitan (1961, p. 61):

> The detrimental effects upon adaptive abilities due to acutely destructive lesions such as intrinsic tumors or cerebral vascular accidents may be more dramatic than the effects of relatively static conditions such as healed head wounds or slowly progressive conditions.

Indirect reference to this uncontrolled variability has been discussed by Yates (1954a) in which he urged the use of comparable experimental groups in the replication of studies in this area. For example, paretics comprised the majority of experimental subjects (*S*s) in the validation of the Hunt-Minnesota Test for Organic Brain Damage (Hunt, 1943), whereas traumatic head injuries were used, for the most part, in the replication of this test (Aita, Armitage, Reitan, & Rabinowitz, 1947). The latter study failed to substantiate the findings reported by Hunt. Paresis generally involves a chronic and irreversible condition of the brain, whereas the effects of traumatic head injury are often transient and reversible (Jasper, Kershman, & Elvidge, 1945).

The only study which has attempted a systematic comparison of types of brain dysfunction was reported recently by Fitzhugh et al. (1961). The classification consisted of *acute* types, in which neurological signs were due to a specific, temporally defined brain lesion; *relatively static* types which consisted of those patients who had either recovered from acute episodes, if any, or who exhibited slowly progressive brain disease without evidence of sudden onset; and *chronic-static* types which were composed of institutionalized patients having longstanding brain dysfunction. Results were in the expected direction of greater impairment for patients who suffered from acute organic damage than for patients having relatively static damage (e.g., posttraumatic concussions, psychomotor epilepsies) or chronic-static damage (e.g., convulsive disorders of the grand mal type). Classification of the independent variables, as such, has a twofold effect. It increases the efficiency of the test(s) by reducing the amount of uncontrolled variance, and further increases our understanding with respect to the behavioral correlates of types of brain dysfunction.

Localization and Lateralization of Brain Lesions

In the last decade there has been much emphasis on the search for specific and localized functions in the brain. The most convincing support of specific brain functions comes from studies of the language process. The left hemisphere has been found to govern language functions almost exclusively (Penfield & Roberts, 1959). Patients with left temporal-parietal lesions have shown consistent deficits on tests which require verbal reasoning and informational skills (Dennerll, 1964; Reitan, 1955, 1962; Satz, 1966; Weinstein & Teuber, 1957). Further, these deficits have occurred in the presence or absence of clinically described language disorder. The right hemisphere, on the other hand, has been shown to have a different pattern of functional organization. Damage to this hemisphere has been associated with perceptual difficulties in manipulating, ordering, and effecting spatial relationships. Such deficits have been measured by nonverbal tasks and have been defined as visuo-constructive functions (Milner, 1954, 1962; Reitan, 1955; Teuber, 1962).

General support for the assumption that the cerebral hemispheres mediate differential functions could have significant value in the design of new tests sensitive to brain dysfunction and localization. However, the lateralization of visuo-constructive functions to the right cerebral cortex has not been entirely confirmed. Costa and Vaughan (1962), employing a series of complex perceptual, motor, and verbal tasks on patients with lateralized cerebral lesions, found

constructional and perceptual deficits for *both* left and right hemispheric cases, although maximal impairment was obtained for the right brain lesions. Maximal deficit for the left brain lesions, however, was in the predicted direction of impaired verbal performance. This deficit was found exclusively in patients with left hemispheric damage. Similar results were obtained in an earlier study by Heilbrun (1956). Employing tests of language and tests of nonverbal visuo-constructive performance for patients with lateralized cerebral lesions, he found identical impairments for *both* lateral brain-lesion groups on the nonverbal tasks, but demonstrated that verbal deficits resulted primarily from left hemispheric damage.

These findings suggest that the organization of verbal processes is strongly lateralized in the left cerebral cortex, particularly with right handers, whereas the organization of nonverbal perceptual and visuo-constructive skills is probably more diffusely represented in both hemispheres, with possibly greater representation in the right cerebral cortex. This position is consistent with physiological data on somesthesis (Semmes, Weinstein, Ghent, & Teuber, 1960), and more recent findings on the disruption of visuo-constructive skills after either left or right hemispheric damage (Arrigoni & De Renzi, 1964).

Nonspecific Effects

There is also evidence to suggest that on more *complex* nonverbal perceptual tasks (e.g., embedded figures), deficits will occur irrespective of locus of lesions, or laterality (Teuber & Weinstein, 1956). The Embedded Figures Test (EFT; Teuber, 1959) requires the rapid detection of "hidden" figures, that is, of complex line drawings concealed by embedding them in interlacing contours. Deficits on this test transcended the area of visual field defects, and revealed similar impairments for men with injuries in any lobe, or in either or both hemispheres. According to Teuber (1959) the effects of brain lesions in man are twofold—specific and general, and vary as a function of the type of test(s) employed. Similar findings have been observed in Lashley's work with the rat. Lashley (1960) demonstrated that habits of the simple conditioned reflex type were dependent only upon the specific sensory areas involved. For more complex tasks, however, such as the multiple-stick problems, deterioration occurred after substantial lesions to any part of the cortex.

Similar nonspecific effects have been found in somesthesis (Teuber, 1959). Again, the kind of symptom obtained (i.e., specific or general) varied as a function of the kind of tests employed. For example, many of the simple classical sensory tests yielded relatively circumscribed deficits, restricted to a body region opposite to a unilateral brain injury when the central sector of the brain was involved. However, performance on a complex tactile task (a modified Sequin-Goddard formboard test), which was constructed to represent a logical and perceptual analogue of the embedded figures task, was essentially nonspecific with respect to area of damage. Impairment was found for all brain-injury groups, irrespective of locus of lesion, or laterality.

These findings (Teuber, 1959) would seem to deserve some consideration in attempts to design tests diagnostically sensitive to brain dysfunction. Selection of a task should require some measure of complex visual or somesthetic performance similar to the "embedded figures" or formboard tasks previously discussed. The fact that these complex measures were sensitive to brain lesions, irrespective of their locus or laterality, offers some promise for further research in this direction. Apparently underscoring the importance of these complex tests is the fact that adequate performance depends on a number of different psychologic functions, each of which is crucial. Hence, any lesion sufficient to affect one of these functions may lead to a significant *general* impairment.

Problem of Test Selection

The preceding discussion has focused on some of the methodological and theoretical problems that have often been ignored in attempts to construct diagnostic tests of brain dysfunction. There are, however,

additional problems that should be considered—namely the choice of the test itself. Although the preceding discussion articulated some of the relevant independent and dependent variables, the problem still remains that brain damage very often leads to changes which are minimal and elusive, and which require very special tasks for their discovery (Teuber, 1959). Standard intelligence tests have notoriously failed to detect deficits in IQ following destruction of significant parts of the human brain (Hebb, 1949; Meyer, 1961; Weinstein & Teuber, 1957; Yates, 1954a). A puzzling finding from these studies is the apparent resiliency of intellectual functions subsequent to brain injury. But is this statement true? For normal persons, vocabulary and information tests are presumed to be the best individual measures of general intelligence. These tests are further assumed to give us the best estimates of a person's level of problem solving outside the laboratory or clinic, in a wide range of situations. According to Hebb (1949, p. 290), the nature of the problem is as follows: "How can they (the tests) also be the ones that show the least effect of brain operation, and the degenerative changes of senescence?"

Hebb reasoned that the apparent resiliency of intellectual functions after brain injury was due to the heavy loading of stored-information experiences on tests such as the Binet and the Army General Classification Test (AGCT). He distinguished between two components of intelligence, one which refers to an innate capacity for acquisition (Intelligence A), and the other which refers to the functioning of the brain in which development has already occurred (Intelligence B). The latter concept involves day-to-day experiential functioning already based on a history of learning. According to Hebb, it is this functioning level that is largely measured in psychometric tests of intelligence. After behavior has been learned or acquired, it then loses its dependency on the underlying neural mechanisms and can persist in spite of damage to this brain tissue. There is, however, a concomitant decrease in the likelihood of new learning as a result of this tissue injury. This suggests that the more sensitive indicator of impairment would involve measurement of Intelligence A, rather than B. Hebb adopts a phase or temporal model for intelligence which predicts that these two components of intelligence are differentially affected in children and adults after brain injury. The hypothesis is advanced that the performance of children on intellectual tasks is more fully determined by Intelligence A, and for adults, by Intelligence B. This hypothesis is consistent with findings on the irreversibility of intellectual impairment in the brain-injured child (Hebb, 1949; Strauss & Lehtinen, 1950). The same theoretical viewpoint might also explain why so many brain-injured adults are able to perform adequately on intellectual tasks involving familiar problems associated with long-established habits.

Hebb's position (1949) on intelligence is quite similar to the one advanced by Halstead (1947). Halstead also distinguished between two components of intelligence, "psychometric" (Intelligence B) and "biologic" (Intelligence A) and reasoned that conventional IQ tests are heavily dependent on "psychometric" components, that is, familiar problems associated with long-established habits. Halstead (1947) suggested that this practice has led to a masking of deficits on conventional IQ tests, which otherwise would reflect significant group differences due to brain injury. It would seem, then, that the desirable features of a diagnostic test for brain damage should include, at least in part, some measure of biologic intelligence (Intelligence A), and should require minimal dependence on previous experience for the solution of the task. It is interesting to note the similarity in this position with Goldstein's (1959) hypothesis that the capacity for learning is significantly impaired in organic brain disorder.

The preceding discussion suggests the feasibility of employing a complex measure of visual or somesthetic performance as a promising dependent variable in the construction of a diagnostic test for brain damage (Teuber, 1959). These tests also seem to meet the requirements advanced by Hebb (1949) and Halstead (1947). They are sufficiently complex measures which require

minimal use of language; they introduce new kinds of learning situations and they require only minimal use of memory or past experience for their solution. Unfortunately, none of these complex measures has yet provided sufficient discriminatory power for diagnostic classification (Teuber, 1959). The experimental, rather than practical diagnostic setting in which these tests have been used, would perhaps account for this shortcoming. Second, there has been no attempt to evaluate the performance of psychiatric patients, particularly schizophrenic *S*s, on these special tests. Third, performance on these tests has been shown to correlate disproportionately with general intelligence. For example, Teuber and Weinstein (1956), employing the EFT, obtained a Pearson *r* of .75 with the AGCT, which suggests that test complexity has been obtained at the expense of an undesirable increase in difficulty. It also suggests the need to partial out the effects of general intelligence on complex visual-perceptual measures.

The Rotation Phenomenon

In the search for a more efficient psychological measure it would be difficult to ignore the experimental and clinical literature dealing with the phenomenon of rotation. This particular visual-motor error tendency was first observed on the Goldstein-Scheerer Cube Test, which requires *S* to reproduce patterns with 1-inch colored cubes (Goldstein & Scheerer, 1941). Some patients (mostly organic) while completing the designs correctly left the pattern in a rotated position, apparently without the awareness of the rotation. Such distortions were tilted at an angle to the target design, sometimes as much as 45 to 90 degrees. This perceptual error has subsequently been observed on several different types of visual-motor tasks (e.g., the Bender-Gestalt, the Graham-Kendall Memory-for-Designs Test (MFDT), Benton's Memory for Designs, and the WAIS Block Design subtest). Rotation errors on these tasks have consistently shown a positive relationship to brain injury and mental retardation (Bender & Teuber, 1948; Griffith & Taylor, 1960a; Pascal & Suttell, 1951). Attempts to use these rotation measures as diagnostic aids, however, have not produced scorable indices of demonstrable validity (Chorost, Spivak, & Levine, 1959). Some of the difficulty is attributable to the infrequent occurrence of rotation errors on the stimulus designs used. The tests, for example, were not constructed to elicit and/or measure this response tendency.

The most systematic attempts to investigate this "perceptual anomaly" have been made by Shapiro (1951, 1952, 1953). In the Block Design Rotation Test (BDRT), which was devised by Shapiro, *S* had to reproduce various stimulus designs presented by the experimenter (*E*) which were composed of either square or diamond figures placed on either square or diamond backgrounds. The task was designed specifically to test the assumed relation between rotation and certain geometric properties of the stimulus designs. Rotation was measured as the number of degrees by which *S*'s design differed in orientation from that of the stimulus design. Shapiro (1953) found that brain-injured *S*s rotated their reproductions significantly more than normal and psychiatric controls. He hypothesized that the greater rotation tendencies in brain-damaged *S*s were due to an increase in cortical inhibition caused by trauma which left the patients peripherally blind to ground determinants of the stimulus. Although Shapiro's findings have been replicated in subsequent studies (Williams, Lubin, Gieseking, & Rubenstein, 1956, 1961), his theoretical interpretation has been rejected. Williams et al. (1956, 1961) experimentally used field reducers in order to block out peripheral cues on this task and found that the organics significantly reduced their rotation errors under this condition while the normals significantly increased their rotations. The authors concluded that whereas the normal *S* probably uses peripheral cues as guides to correct orientation of his designs, the brain-injured *S* may be confused and distracted by them. This might explain the organic's tendency to become "stimulus-bound" on complex perceptual tasks (Goldstein, 1959). The stimulus boundedness would result from a failure to avoid the distracting peripheral cues which would consequently reinforce *S*'s attention to more concrete attributes of the

figure. These conclusions are in agreement with those advanced by Strauss and Lehtinen (1950) who state that brain-damaged *S*s are easily distracted by stimuli, and that appropriate retraining methods should involve gradual reduction of peripheral stimulation during the relearning process.

Rationale

The preceding studies suggest the usefulness of the rotation effect as a diagnostic measure of brain damage. The BDRT, proposed by Shapiro (1953), has several of the attributes of a complex perceptual task which Teuber and co-workers have found sensitive to lesions, irrespective of locus or laterality in the brain. Like the EFT, the BDRT shares many of the attributes of a "biologic" measure of intelligence. Unlike the EFT, however, the BDRT has the advantage of having been designed in a clinical setting as a diagnostic measure for brain disorder. The BDRT is not without certain limitations, however. First, its diagnostic efficiency has never been greater than 75% correct classification. This has been due primarily to its failure to classify correctly *S*s who fall in the dull normal and lower ranges of intelligence (Williams et al., 1956, 1961; Yates, 1954b). Second, the BDRT, like most diagnostic tests, was constructed as a single variable instrument in which the effects of age and IQ were not systematically partialled out. Third, the efficiency of the BDRT has not been studied under conditions of varying base rate probabilities for different diagnostic populations. A fourth criticism relates to its laborious and expensive scoring procedure; each design must be photographed after each reproduction, and only later subjected to detailed measurement. Finally, it is doubtful that the stimulus properties in the BDRT provide the optimal conditions for eliciting the rotation effect. For example, the mean rotation score reported for organics has been small (approximately 7 degrees). Even then, logarithmic transformations have been necessary to correct for the extreme negative skewness of the rotations obtained (Yates, 1954b).

With the assumption that the rotation effect is a potentially useful procedure for the detection of organic brain dysfunction, the following methods of measurement and analysis are proposed for this study: (*a*) construction of a new test of the rotation effect, involving multiple measures, in which the *S* is required to rotate a certain number of degrees and in a particular direction on each of the stimulus problems; (*b*) the application of multivariate discriminant analysis for the prediction of discrete criterion groups with this instrument; and (*c*) the application of base rate and cost efficiency analyses to determine the utility and predictive validity of this multivariate instrument.

METHOD

Nature of Task

The psychological measure proposed was a 44-item visual-motor task designed for optimal measurement of the rotation effect. The Block Rotation Test (BRT) involved a series of newly constructed block designs employing the WAIS blocks (solid red and white colors only). Sample designs are shown in Figures 1 and 2. The task was composed of two parts. Part A (Figure 1) consisted of 15 designs which were presented in a vertical or horizontal position to *S*. Part B (Figure 2) consisted of seven designs which were presented at various angles to the vertical-horizontal axis.

The *E* sat across the table and in front of *S*, and constructed each stimulus design with the blocks. The *S* then reproduced each of the stimulus designs (given additional blocks) as they would look if rotated 90 degrees, either to the left or to the right, in randomized order. The *S* was not allowed to turn the stimulus design and was required to manipulate only *one block at a time*. The blocks presented had both color (red and white) and design properties, but the designs were made simpler than those of the WAIS Block Design subtest in order to limit the general effect of psychometric intelligence. Such a simplification was achieved by using only solid colors for any one block and by reducing the complexity of the total design. It was further assumed that the difficulty of the designs might be reduced by eliminating the element of symbolic representation (i.e., design cards) and restricting the stimulus designs to actual blocks.

The unique feature of this task is that *S* was required to rotate on every stimulus design. Furthermore, the task required a different kind of performance from *S*. He was not required to reproduce block designs as presented in abstract stimulus cards; and he found little, if any, opportunity to utilize past experience in the solution of the task. In short, he was no longer required merely to reproduce, recognize, or discriminate various complex stimulus materials. He was confronted with a new test situation in which the solution was not

embedded in his stimulus field. In terms of figure-ground properties of the task, the *S* was required to shift or alternate constantly between figure and ground perceptions. The figure was defined as the stimulus design which *E* constructed and placed in front of *S*. The ground determinants were defined as all possible degrees of rotation from the given stimulus design or figure. Only one particular ground-spatial relationship, however, was correct for each stimulus figure, that is, a 90-degree rotation either to the left or to the right from the stimulus (depending on *E*'s instructions). What was originally figure (i.e., the stimulus design) had to be converted into a new figure which was part of the spatial background for that percept. The difficulty arose from the fact that the ground determinants were neither present, embedded, or concealed; and they were not in the *S*'s stimulus field. As such, they required the *S* to resort to "inner" inferences, regarding the appropriate background rotations, which were assumed to involve more central brain processes. In other words, it was felt that the absence of the background visual cues would displace the reliance on sensory field information to more centrally determined visual-perceptual processes in the brain.

Administration and Scoring[3]

Detailed instructions and scoring method are presented in the appendix of the author's dissertation (Satz, 1963, pp. 178–200). Testing was performed on a table, approximately 26 inches square. The *E* sat directly across from *S* throughout the testing procedure and constructed each block design one at a time from a stack of 3 × 5 design cards which were placed in front of *E*. There were two sets of design cards: one for Part A, which consisted of 15 printed cards plus 2 printed example cards, and one set for Part B, which consisted of 7 printed design cards plus 1 printed example card. Selected designs, for each part, are presented in Figures 1 and 2. Each card also depicted the order of directional turn for the particular design (i.e., left or right). To the right of *E* was a single 3 × 5 scoring card which listed by number the 15 block designs on Part A and the 7 block designs on Part B. This recording procedure allowed *E* to analyze errors with respect to individual designs for both left- and right-turn rotations on both parts of the test. The *S*'s name, age, and education were also recorded on this card.

Part A of the test was administered first and was preceded by two example designs which were not scored. The first example was a horizontal design composed of two solid blocks (one red and one white) which were adjacent to each other. The second design was a vertical design composed of two solid blocks (one red and one white) which were adjacent to each other.

[3] All testing on the initial standardization was performed by the present author. In the subsequent cross-validation studies, however, all testing was performed by other laboratory personnel.

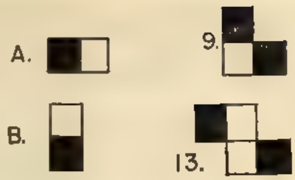

FIG. 1. Sample designs, Part A.

To briefly summarize the procedure, *S* was first shown how the stimulus design (Example A) would look if it were turned to the left or to the right 90 degrees. The *E* did this by turning the stimulus design 90 degrees in both directions. The *S* was then given the opportunity to practice these 90-degree quarter turns on Example A. When *S* was able to turn this first example design to a criterion of three successes for each direction, he was then required to discontinue turning and to attempt to reproduce this design (given additional blocks) as it would look if turned 90 degrees to the left and to the right. When *S* completed this design correctly, for both turns, administration proceeded to the second example involving the vertical design. The *S* was requested not to turn this design, but to picture how it would look if turned 90 degrees in the desired direction and then to build this visual image. The *S* was permitted to rotate the stimulus design only if he failed to make the appropriate rotation. When *S* completed this design correctly, for both turns, testing proper began.

Part B of the test was preceded by only one example design which was made at a 45-degree angle to the vertical axis. The design was also composed of two solid blocks (one red and one white) which were adjacent to each other. Like the first example on Part A, *S* was first shown how the stimulus design (Example A) would look if it were turned to the left or to the right 90 degrees. The *E* did this by again turning the stimulus design 90 degrees in both directions. The *S* was then given the opportunity to practice these quarter turns on Example A. The *S* was told, on this part of the test, that each design would end up at an angle due to the nature of the 90-degree turn. When *S* was able to turn this example design to a criterion of three successes for each direction, he was then required to discontinue turning and to attempt to reproduce this design (given additional blocks) as it would

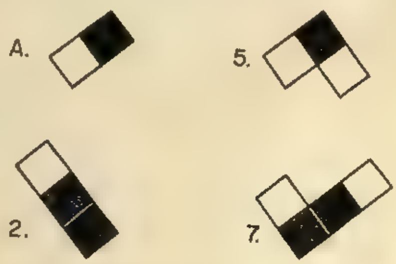

FIG. 2. Sample designs, Part B.

look if turned 90 degrees to the left and to the right. When *S* completed this design correctly, for both turns, testing proper began. Both parts of the test were scored only for error responses which are discussed below.

Part A errors. Part A consisted of 15 designs with either horizontal or vertical axes. The first scorable design was composed of two blocks similar to the two example designs on this part of the test. Seven of the remaining designs were composed of three blocks, and seven were composed of four blocks. Each of these designs had to be rotated 90 degrees in both directions. An error was defined as any design made by *S* which was not a reproduction of the stimulus design as it would look if turned 90 degrees in the requested direction. It was recorded by a checkmark (√) on *E*'s 3 × 5 scoring card for the particular design and turns under Part A. The maximum number of scorable errors (√) was 30 for this part.

Part B errors. Part B consisted of seven angular designs (i.e., the axes were at a 45-degree angle to the vertical). The first scorable design was composed of two blocks similar to the example design for this part of the test. Four of the remaining designs were composed of three blocks and two were composed of four blocks. The only principle of symmetry which guided this choice of alternate design patterns, for each part of the test, came from empirical observation during the pilot stages of the study. Each of the designs on Part B were also scored for left and right turns. An error was again defined as any design made by *S* which was not a reproduction of the stimulus design *as* it would look if turned 90 degrees in the requested direction. It was also recorded by a checkmark (√) on *E*'s 3 × 5 scoring card for the particular design and turn under Part B. The maximum number of scorable errors (√) for this part was 14.

Total errors. Errors were summed separately for Part A and for Part B of the test, and they were also combined, giving a maximum possible Total error score of 30 + 14 = 44.

Types of errors were also scored. Under this classification there were three distinct variables which are listed as follows:

Duplication errors. The Duplication error was defined as any design made by *S* which merely reproduced *E*'s stimulus design. It was recorded by the symbol DE on *E*'s scoring card and was classified both as an error and as a type of error. The symbol DE, therefore, indicated a twofold classification: an error response (√) and the type of error involved.

Angulation errors. The Angulation error was defined as any design made by *S* which was angulated or tilted on Part A, or was horizontal or vertical on Part B. This type of error score was a logical consequence of the fact that Part A required horizontal-vertical designs for the correct rotated position, and Part B required diagonal designs for the correct rotated position. These errors were recorded by the symbol AE on *E*'s scoring card, and were also classified as an error response (√) and as a type of error.

Time errors. The Time error was defined as a failure to complete the rotated design within a 65-second time limit. In the administration of the test, *E* allowed a 5-second interval on all designs between giving the request for a left or right turn, and placing the blocks before *S*. The *S* was then allowed 60 seconds to make the appropriate design rotation. The procedure of employing a 5-second interval, before placing the blocks before *S*, was used in an attempt to provide *S* with an opportunity to visualize the rotated stimulus design before manipulating the blocks; it was felt that this procedure would reinforce emphasis on the perceptual rotation and also reduce trial and error behavior. This error was recorded by the symbol TE on *E*'s scoring card, and also indicated an error response (√) and the type of error involved.

There were, in summary, six measures of error which were not all independent: (*a*) Part A errors, (*b*) Part B errors, (*c*) Total errors, (*d*) Duplication errors, (*e*) Angulation errors, and (*f*) Time errors. An error, in general, was defined, once again, as any design made by *S* which failed to reproduce the stimulus design as it would look if turned 90 degrees in the requested direction. This operational definition was both a necessary and sufficient condition for all errors, regardless of whether they were classified by type or not. For types of errors, however, this definition was a necessary, but *not* sufficient condition. For example, a design which deviated from what the stimulus design would look like, if rotated 90 degrees in the appropriate direction, was always an error (√), but for this design to constitute a particular type of error, it had to meet the specified requirements already defined for this class of error. Whether it fulfilled these requirements or not, it was still an error (√). Since it was possible for a person to make exclusively Duplication errors, or Angulation errors, or Time errors on both parts of the test, the maximum number of errors possible for each type of error was equal to the maximum number of errors obtainable on both Part A (i.e., 30) and Part B (i.e., 14). For example, if a person made 10 Duplication errors on Part A and 3 Duplication errors on Part B, then his error scores would be classified under at least the following four variables: Part A errors ($N \geqq 10$); Part B errors ($N \geqq 3$); Total errors ($N \geqq 13$), Duplication errors ($N = 13$). The same would be true for Angulation and Time errors.

Additional Test and Nontest Variables

Along with the six error variables measured by the BRT, five additional test and nontest variables were further included in this study. The two nontest variables were age and education. Age was included because of its relationship to both "psychometric" intelligence (Wechsler, 1958) and "biologic" intelligence (Reitan, 1956). Educational level was included because of its relationship to socioeconomic level and general intelligence (Griffith & Taylor, 1960b). The three test variables were the Abbreviation of the WAIS for Clinical

Use (Mogel & Satz, 1963; Satz & Mogel, 1962), the Trail Making Test (TMT) (Armitage, 1946; Reitan, 1958), and the MFDT (Graham & Kendall, 1960).[4] Only the Performance IQ Scale of the Abbreviated WAIS was administered; this was to insure a nonverbal psychometric IQ control for the BRT. Pearson r correlations between the Abbreviated and original WAIS forms have been reported as follows (Satz & Mogel, 1962): Verbal IQ = .99; Performance IQ = .97; Full Scale IQ = .99. Correlation coefficients of this magnitude were found regardless of intellectual level or diagnostic classification. Subsequent research has confirmed these findings (Estes, 1963; Pauker, 1963; Watson, 1966). The TMT and MFDT are standardized tests for brain damage and were included as a comparative means of evaluating the efficiency of the BRT.

Method of Analysis

The discriminant function (Fisher, 1936) was used in the analysis of the data. This statistical technique was devised for the problem of maximally differentiating discrete criterion groups when multiple measurements are involved. The technique is essentially a multiple regression problem except for the discontinuous distribution on the criterion variables. The expression of this function is given by the following linear equation:

$$Z = \lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + \cdots + \lambda_n x_n \quad (1)$$

in which Z is the composite or compound predictor score based on the individual scores on each of the variables or tests employed ($x_1, x_2, x_3, \cdots x_n$), and on the respective weights (i.e., lambdas) assigned to each of these scores ($\lambda_1, \lambda_2, \lambda_3, \cdots \lambda_n$). The task for a problem involving only two criterion groups, when multiple measures are involved, is to determine the optimal weights for these variables which will make the difference between the composite Z measures on both criterion groups as large as possible; or, in other words, to find values for the lambda coefficients which will maximize the difference between the composite means of the two criterion groups involved (Garrett, 1943; Goulden, 1956).

Subjects

The standardization sample (N = 122) consisted of four groups of adult males: normals (N = 20), neurotics (N = 20), schizophrenics (N = 23), and organics (N = 59). The organic brain disorders were selected from two different Veterans Administration Hospitals; one subgroup (N = 33) included samples from the Acute-Intensive Treatment Service at the Veterans Administration Hospital, Lexington, Kentucky; the other subgroup (N = 26) was composed of samples from the neurological wards at Hines Veterans Administration Hospital, Chicago, Illinois. The latter subgroup was included in order to provide more representative cases of neurological brain involvement. All diagnoses for the organic patients were based upon decisions of the hospital medical staffs. These decisions were based upon detailed medical history, electroencephalography, neurological examination, and when further classification was needed, angiography and pneumography.

The organic patients were classified by *type* of brain involvement (N = 59), and by *area* of brain involvement when objectively possible (N = 35). These classifications were employed in an attempt to account for some of the variability in performance among brain-injured patients.[5]

Classification by type of brain damage was made according to the criteria advanced by Fitzhugh et al. (1961, p. 61). One group, *acute*, was composed of patients (N = 23) who had acute neurological illnesses and whose neurological symptoms were present at the time of psychological testing. These patients had experienced a specific temporally defined episode, during which their current neurological findings had arisen, or had developed a rapidly progressive brain disease with steady progression of neurological signs. A second group, *relatively static*, was composed of patients (N = 24) who had either recovered from acute neurological signs, or who had slowly progressive brain disease without evidence of sudden onset. Among this group, the patients with sudden onset of brain dysfunction (e.g., brain trauma) had with the passage of time recovered from acute neurological deficits, suggesting a reorganization of brain function and a relatively static condition of the brain. The third group, *chronic-static*, was composed of patients (N = 10) with chronic, long-standing brain dysfunction. The patients in this group consisted largely of prefrontal lobectomies and chronic epileptics. Diagnoses of the patients within each group according to type of brain involvement are presented in Table 4.

Classification by area of brain involvement was as follows: prefrontal lesions (N = 10), which included primarily cases of prefrontal lobectomy; left temporoparietal lesions (N = 15), which included for the most part cerebral vascular accidents (CVA) and brain tumors; and right temporoparietal lesions (N = 10), which for the most part included CVA and brain tumor cases.

The neurotic, schizophrenic, and normal *Ss* were all selected from the Veterans Administration Hospital, Lexington, Kentucky. The neurotic and schizophrenic patients were selected from the Acute-Intensive Treatment Service of the hospital. The normal group was also selected from

[4] The Trail Making and Memory-for-Designs Tests were administered by the present author, but the scoring was performed by Walter Lindley, a Clinical Psychology Trainee at the Veterans Administration Hospital, Lexington, Kentucky. Scoring was performed without knowledge of the *Ss'* diagnosis, age, or Performance IQ.

[5] In the final restandardization (DF III, 1966) the organic patients were also grouped by classification of lesion (e.g., vascular, neoplastic, convulsive).

TABLE 1

MEANS AND DIFFERENCES BETWEEN CRITERION GROUPS ON THE PREDICTOR VARIABLES
STANDARDIZATION GROUP (DF I)

Variables	Nonorganic (N = 63)		Organic (N = 59)		t
	Mean	SD	Mean	SD	
Part A	3.08	2.56	11.83	7.09	8.95*
Part B	2.73	1.86	7.78	3.07	10.88*
Total	5.81	3.85	19.61	9.30	10.58*
Duplication	.52	.98	4.46	4.56	6.49*
Angulation	.22	.52	2.29	3.13	4.99*
Time	.32	.67	.97	1.63	2.84*
Age	36.22	8.09	39.66	8.74	2.26*
PIQ	95.98	10.79	91.73	11.07	2.15*

* $p \leq .05$.

this service and consisted of male psychiatric nursing assistants. The diagnoses for the neurotic and schizophrenic patients were contingent on the final decisions of the medical-psychiatric staffs. In no case was a neurotic or schizophrenic patient included if he had a past history of head injury or neurological disease. The same criteria applied for selection of the normal *S*s.

Hypotheses

The following hypotheses were addressed to possible dimensions within the concept of brain disorder, that is, *type* of damage, *area* of damage, and *classification* of disease.

Hypothesis 1. Changes in test behavior after cerebral damage will vary as a function of the type of damage involved. In line with Fitzhugh et al. (1961), *acute* types of brain damage will reveal more significant impairment in test performance than either *relatively-static* or *chronic-static* brain-injury cases. Discriminant function composite scores were used in this analysis.

Hypothesis 2. Changes in test performance after cerebral damage will show nonspecific and generalized effects on a test of complex perceptual functioning such as the BRT. Specifically the hypothesis predicts that performance will be unrelated to either locus or laterality of brain involvement, due to the complex nature of the task employed. This hypothesis (Teuber, 1959) is in contrast to the alternative hypothesis that would predict maximal impairment, on visuo-constructive tasks, for right hemispheric damage (Milner, 1962; Reitan, 1962). Discriminant function composite scores were used in this analysis.

Hypothesis 3. Changes in test behavior will vary as a function of *classification* of disease. This hypothesis is addressed to the possible differential effects between structural lesions (e.g., neoplastic and vascular conditions) and nonstructural lesions (e.g., convulsive disorders and toxic drug conditions). Discriminant function scores, based on the final restandardization group (DF III), were used in this analysis.

RESULTS

Table 1 presents the mean differences between the *combined control* (Group I) and *organics* (Group II) on the eight predictor variables which showed the best discrimination between criterion groups.[6] The variables included the six BRT variables along with Age and Performance IQ. Inspection of this table reveals significant differences between criterion groups on each variable. The differences were in the direction of higher scores (i.e., errors) for the organics on all but the Performance IQ variable. Performance IQ was lower for the brain-injured group ($p < .05$). The three variables which were excluded from further analysis were the TMT, MFDT, and education. Although the TMT correctly classified 85% of the brain-injured *S*s (Group II), it also misclassified 73% of the combined controls (Group I). The MFDT, on the other hand, correctly classified only 49% of the brain-injured *S*s (Group II) and misclassified 42% of the psychiatric and normal controls (Group I). Education also showed a high overall rate of misclassification (valid positive = 46%; false negatives = 40%). In fact, the mean and standard deviation scores for education were quite similar for

[6] The reader is referred to the author's doctoral dissertation (Satz, 1963) for detailed information concerning the statistical selection of predictor variables, and the reasons for combining the normal, neurotic, and schizophrenic groups. The decision to combine the three nonorganic groups into one criterion group was based essentially on the failure to differentiate these groups on the predictor variables.

TABLE 2

INTERCORRELATIONS FOR TOTAL SAMPLE[a] OF EIGHT PREDICTOR VARIABLES

Variable	2 Part B error	3 Total error	4 Duplication error	5 Angulation error	6 Time error	7 Age	8 Performance IQ
1. Part A error	.769	.973	.740	.378	.059	.321	−.416
2. Part B error		.896	.602	.499	.067	.293	−.354
3. Total error			.731	.443	.065	.329	−.417
4. Duplication error				.195	−.008	.305	−.090
5. Angulation error					.008	.153	−.259
6. Time error						.183	−.137
7. Age							.031

Note.—Correlation coefficients of .178 and .233 are significant at the .05 and .01 levels, respectively.
[a] N = 122 (20 normals, 20 neurotics, 23 schizophrenics, and 59 organics).

each of the criterion groups (Combined controls, $\bar{X}$ = 10.17, SD = 2.77; Organics, $\bar{X}$ = 10.31, SD = 3.20).

Table 2 shows the Pearson product-moment correlation coefficients for the eight variables on the initial standardization sample (N = 122). Although there were a number of significant correlations (n = 19), only a few were large enough to account for a sizable amount of the variance. The highest correlation obtained was between Part A and Total errors (r = .97); the inflation of this correlation coefficient was due in part to the confounding effects of Part A which was included in the summation of Total errors. The relationship of Age and Performance IQ with the other measures in this correlation matrix was of particular interest. Both variables correlated significantly with the three best individual predictors on the BRT. Performance IQ was inversely related to Part A errors (r = −.42), Part B errors (r = −.35) and Total errors (r = −.42); Age was positively related to Part A errors (r = .32), Part B errors (r = .29) and Total errors (r = .33).

The Discriminant Analysis (Discriminant Function I, 1963)

The discriminant function analysis was performed on the eight variables presented in Tables 1 and 2. The criterion groups were classified dichotomously, that is, nonorganic (Group I) or organic (Group II). By computing the within-group sums of squares and cross products for all combinations of eight variables, a set of eight simultaneous equations was obtained.[7] The solution of these equations yielded the following lambda values for each variable: Part A (λ_1 = −5.3141), Part B (λ_2 = 3.5168), Total (λ_3 = 12.4600), Duplication (λ_4 = −1.000), Angulation (λ_5 = 7.000), Time (λ_6 = 25.3278), Age (λ_7 = −2.0666), and Performance IQ (λ_8 = 3.9282).

The data were next analyzed to determine the mean composite discriminant score for each criterion group, where the organic group was defined as Population A and the nonorganic group, Population B. For example, the mean composite discriminant score for the organic group would have the following expression: $Z_A = \lambda_1 X_{A_1}, + \lambda_2 X_{A_2} + \cdots + \lambda_8 X_{A_8}$, in which Z_A represents the composite discriminant score based on the mean scores of each variable for the organics ($X_{A_1}, \cdots X_{A_8}$), and the corresponding weights assigned to each of the respective variables ($\lambda_1, \cdots \lambda_8$). The results were as follows: Z_A = 523.24, Z_B = 376.86.

In order to reach a decision on any individual's composite score, the following strategy was adopted to determine the optimal cut-off score:

If $Z_i \geqq \dfrac{Z_A + Z_B}{2}$,

then predict Population A (organic).

If $Z_i < \dfrac{Z_A + Z_B}{2}$,

then predict Population B (nonorganic).

[7] The data were originally computed on a Monroe calculator and later reanalyzed on an IBM 709 computer at the University of Kentucky Computing Center.

TABLE 3

PREDICTIVE CLASSIFICATIONS BY USE OF DISCRIMINANT FUNCTION I[a]

Interval composite scores	Normals $N = 20$	Neurotics $N = 20$	Schizo-phrenics $N = 23$	Organics $N = 59$
675–699				1
650–674				2
625–649				2
600–624				5
575–599				6
550–574				8
525–549				6
500–524		1		4
475–499				7
450–474			1	7
425–449	3	1	4	5
400–424	3	1	3	3
375–399	5	2	5	1
350–374	4	7	5	1
325–349	4	3	2	1
300–324	1	5	3	

[a] Composite cut-off score: $Z \geqq 450.05$; overall hits = 89%, valid positives = 81%, false positives = 3%.

The ratio, $(Z_A + Z_B)/2$, yielded the composite value of $Z = 450.05$ as the optimal cut-off score for this linear prediction equation.

The discriminant function was next analyzed to test the difference between the composite Z scores for the two criterion groups. This analysis of variance is essentially a test of significance of the discriminant function. The results showed significant differentiation between criterion groups on the composite Z scores ($F = 29.79$, $p < .001$).

Predictions were then made on each individual in the study. This was done by computing the composite discriminant score (Z) for each S and predicting "organic" if his composite score was $Z \geqq 450.05$, and "nonorganic" if his composite score was $Z < 450.05$. The results of this decision policy are presented in Table 3. Inspection of this table reveals that only two nonorganic Ss were misclassified, giving a false positive rate of $p_2 = .03$; the misclassification rate for the organic group, however, was larger, and yielded a valid positive rate of only $p_1 = .81$. In spite of the failure to detect several brain-injured cases, the discriminant function still correctly classified 89% of the total standardization sample ($N = 122$).

Analysis of Types of Brain Damage

Table 4 presents the diagnostic distributions within each of the three types of brain-lesion groups. The results of this analysis are reported in Table 5. The composite predictor means of the respective groups were as follows: *acute* ($Z = 551.77$), *relatively static* ($Z = 489.02$), and *chronic-static* ($Z = 532.78$). It is interesting to note that each of these composite Z scores was above the minimal score critical for organic brain disorder ($Z \geqq 450.05$), although the relatively static brain-lesion group did tend to converge closer to this cutting line. The analysis of variance of these composite predictor

TABLE 4

DISTRIBUTION OF SUBJECTS ACCORDING TO TYPE OF BRAIN DAMAGE

Acute ($N = 23$)		Relatively static ($N = 24$)		Chronic-static ($N = 12$)	
Cerebral vascular accident	13	CBS,[a] brain trauma	10	Bilateral frontal lobectomy	6
ABS,[b] drug intoxication	1	Post-traumatic concussion	1	Unilateral frontal lobotomy	1
Post pneumococcal meningitis	1	Cerebral arteriosclerosis	2	Convulsive disorder, grand-mal	14
CBS, brain tumor	6	Psychomotor epilepsy	2	Psychomotor epilepsy	1
Huntington's Chorea	1	CBS, convulsive disorder	6		
Encephalitis	1	CNS lues, meningoencephalitic	1		
		CBS, drug intoxication	1		
		Parkinson's disease	1		

[a] CBS = Chronic Brain Syndrome.
[b] ABS = Acute Brain Syndrome.

TABLE 5
COMPOSITE PREDICTOR MEANS FROM DISCRIMINANT FUNCTION I BY TYPE OF BRAIN INVOLVEMENT

Type of damage	N	Composite mean
Acute	23	551.77
Relatively static	24	489.02
Chronic-static	12	532.78

Analysis of variance of composite predictor score

Source of variation	ss	df	MS	F	p
Between groups	113.00	2	56.50	4.15	<.05
Within groups	763.00	56	13.63		
Total	876.00				

Differences between group composite means

	Relatively static	Chronic-static
Acute	−62.75**	−18.99
Relatively static		43.76

** $p < .01$.

scores (Table 5) revealed an overall separation between groups ($F = 4.15$, $p < .05$). The only difference between group means occurred between the acute and relatively static brain lesions ($t = 2.83$, $p < .01$), with the acute lesions showing greater impairment. There was, however, no significant difference between the acute and chronic-static groups, although the trend was in the direction predicted. Further analysis revealed that 11 of the 12 *S*s within the chronic-static group (91%) were correctly classified as organic. In view of past difficulties in detecting cases of prefrontal damage, the latter finding is of some interest. The hit rate (H_T) for the acute group was also 91%, with 21 of the 23 *S*s being correctly classified. It is impossible to compare these repsective hit rates with Fitzhugh et al. (1961), because this information was not reported in their study.

An unexpected finding was the fact that 8 of the 11 diagnostic misclassifications for the total organic sample (i.e., 73%) were represented within the relatively static group. Apparently the higher false negative rate ($1 - p_1 = .19$) for the discriminant function was due primarily to the inclusion of relatively static brain-lesion cases, which in turn reduced the valid positive rate ($p_1 = .81$). In spite of this limitation, however, the discriminant function still classified correctly 67% of the relatively static cases.

The results, in summary, did support the general hypothesis that the type of brain lesion is a meaningful variable within the concept of "brain damage." The specific hypothesis relating to the direction of differences between types, however, was only partially confirmed.

Analysis of Generalized Effects

This analysis was performed on 35 brain-injured *S*s who were classified according to locus and laterality of brain involvement. The results are reported in Table 6. The composite predictor means for each group were as follows: left temporoparietal damage ($Z = 546.87$), right temporoparietal damage ($Z = 543.30$), and frontal lobe damage ($Z = 534.70$). The mean composite Z scores for each group were well above the critical cutting line for organic brain disorder ($Z \geq 450.05$). Analysis of variance on the composite scores, however, failed to show any significant overall difference between groups. This finding therefore supports the hy-

TABLE 6
COMPOSITE PREDICTOR MEANS FROM DISCRIMINANT FUNCTION I BY AREA OF BRAIN INVOLVEMENT

Area of damage	N	Composite mean
Left temporo-parietal	15	546.87
Right temporo-parietal	10	543.30
Frontal	10	534.70

Analysis of variance of composite predictor scores

Source of variation	ss	df	MS	F	p
Between groups	900.24	2	450.12	.0975	ns
Within groups	147799.93	32	4618.75		
Total	148700.17				

pothesis of nonspecific effects with this test instrument.

Test-Retest Reliability

In order to evaluate the reliability of the test, the BRT was readministered to a random sample of *S*s ($N = 18$) from the original standardization group ($N = 122$). The Performance IQ Scale of the WAIS (Satz & Mogel, 1962) was also readministered to this sample. The problem was analyzed in two ways: (*a*) in terms of the Pearson product-moment correlations on each test variable between testings and (*b*) in terms of the classification changes on the composite Z scores after retesting. The retest composite Z scores were computed to provide some general measure of classification reliability. Roughly 4 weeks intervened between the test and retest sessions for each *S*.

The Pearson correlation coefficients for each variable were as follows: Part A errors, $r = .89$; Part B errors, $r = .85$; Total errors, $r = .91$; Duplication errors, $r = .81$; Angulation errors, $r = .75$; Time errors, $r = .67$; and Performance IQ, $r = .89$. The correlation coefficients between test and retest sessions were significant for each of the variables ($p < .01$), although three of the variables (Duplication, Angulation, and Time errors) were more subject to change during this test-retest interval. With regard to classification changes on the composite Z scores, only one *S* showed a change in predictive classification after retesting. This involved a schizophrenic patient who was originally misclassified on the first testing, but who was correctly classified on retesting. All other *S*s were predicted within the same criterion group on both testings, regardless of diagnosis. These results, in summary, suggest adequate demonstration of reliability for the individual variables and the composite predictions.

Cross-Validation Studies (1964–1966)

The following studies were undertaken to examine the predictive validity of the BRT in a new diagnostic population, and to control for the possible examiner bias in the original standardization (DF I, 1963). The studies are as follows: (*a*) cross validation of DF I on a new sample of brain-lesion and control *S*s (1964); (*b*) restandardization (DF II) based on the original standardization group (DF I) and the 1964 cross-validation sample; (*c*) cross validation of the DF II on an additional sample of brain-lesion and control *S*s (1965); and (*d*) restandardization (DF III) based on the second standardization groups (DF II) and the 1965 cross-validation sample (1966).

The purpose of these separate validational analyses was twofold: (*a*) to obtain greater representation in the criterion groups in order to stabilize the lambda weightings and (*b*) to determine when adequate predictive validity had been demonstrated.

The *S*s were all selected from the Inpatient and Outpatient Services of the Teaching Hospital at the University of Florida College of Medicine, Gainesville, Florida. All *S*s, from both criterion groups, were given thorough neurological evaluations by the

staff of the Department of Neurology. Several of the neurological patients were also evaluated by the Division of Neurosurgery. Classification of *S*s to the organic criterion group was again based upon detailed medical history, neurological examination, electroencephalography, brain scans, skull films, and when necessary, arteriography, angiography, and pneumography. Classification of *S*s to the nonorganic criterion group was based upon at least a negative medical history and neurological examination, and frequently upon negative EEG and skull-film reports. In roughly 25% of the cases, angiography and/or pneumography was carried out due to the nature of the presenting symptoms.

The nonorganic *S*s were selected from this population in order to provide a more realistic approximation of the typical inpatient medical setting in which patients are likely to present a wide variety of "apparent" neurological disorders. This selection procedure also provided a more detailed examination of the *S*s assigned to the nonorganic criterion group. Detailed compositions of both groups are presented in the final restandardization (DF III).

The BRT and WAIS were administered by the staff and trainees of the Clinical Neuropsychology Laboratory, with the exception of the present author.[8] All testing was done "blind" without reference to hospital or referral charts. This procedure is routinely followed in psychological evaluations in the laboratory.

Discriminant function I: Cross validation. This investigation was made on 100 consecutive referrals to the laboratory during 1964. The sample consisted of 48 brain-lesion cases (Group I) and 52 psychiatric, general medical, and normal *S*s (Group II). Group I was represented equally by vascular, neoplastic, traumatic, and convulsive disorders. Group II was largely represented by psychiatric and general medical cases. Classification was made on the basis of the lambda coefficients derived from DF I and each *S*'s score on the eight predictor variables in the present sample. The original cutting line of $Z \geqq 450.05$ was again used. These results are presented in Table 7. Inspection of this table shows that although the instrument correctly classified 75% of the total sample ($N = 100$), there was a 14% shrinkage in overall hits compared to the original standarization (DF I). This is not too surprising when one considers the fact that a new population was sampled, different examiners were used, and further, that the original standardization was based on a relatively small number of *S*s ($N = 122$). The main source of error classification occurred with Group II in which 29% of the nonorganics were falsely classified. This represents a 26% increase in the false positive rate over the original standardization (DF I). Of greater interest, however, is the striking similarity in the valid positive rate between the two studies. The original standardization function correctly classified 81% of the organics, and 79% of the organics in the cross-validation sample. Further, 5 of the 10 false negative errors were convulsive disorders in which the only criterion evidence was EEG

TABLE 7

Cross Validation of Discriminant Function I (1964 Sample)[a] ($N = 100$)

Interval composite scores	Nonorganics ($N = 52$)	Organics ($N = 48$)
675–699		11
650–674		1
625–649	1	2
600–624		2
575–599	1	1
550–574		1
525–549	2	4
500–524	4	6
475–499	4	8
450–474	3	2
425–449	15	2
400–424	5	2
375–399	6	3
350–374	9	1
325–349	1	2
300–324	1	

[a] Composite cut-off score: $Z \geqq 450.05$; overall hits = 75%, valid positives = 79%, false positives = 29%.

[8] The author is deeply grateful to Eileen Fennell, Research Assistant, who was responsible for supervising the administration of the BRT to predoctoral students and who handled the detailed follow-up classification of patients during the cross-validation analyses.

TABLE 8
MEANS AND DIFFERENCES BETWEEN CRITERION GROUPS ON THE PREDICTOR VARIABLES RESTANDARDIZATION GROUP (DF II)

Variables	Nonorganics ($N = 117$)	Organics ($N = 105$)	t
Part A	3.52	11.27	9.26*
Part B	2.99	7.40	9.55*
Total	6.51	18.67	9.50*
Duplication	.66	3.64	4.11*
Angulation	.62	2.91	3.42*
Time	.38	1.45	2.40*
Age	34.52	41.63	2.08*
Performance IQ	99.05	94.80	2.18*

* All t values significant at $p \leqq .05$.

abnormality and a history of fits, that is, no structural damage was evidenced.

On the basis of these findings it was decided to combine the data of this sample with the original standardization sample and to compute a new discriminant function based on more representative cases.

Discriminant function II: Restandardization (1964). Table 8 presents the means and differences between the organic ($N = 105$) and nonorganic ($N = 117$) groups on each of the discriminant predictor variables for the original standardization group (DF I) and the 1964 cross-validation sample combined.[9] Inspection of this table reveals significant group differences on each variable. Although the neurological group was both older ($t = 2.08, p < .05$) and less intelligent ($t = 2.18, p < .05$), their mean level of intelligence was within the Normal Range ($\bar{X} = 94.80$) and their mean age was not high ($\bar{X} = 41.63$). In other words, Group I was not heavily composed of deteriorated organic patients which, if so, would tend to spuriously inflate the valid positive rate.

The discriminant function analysis was computed on an IBM 709 Computer with the use of the UCLA BIMED program (No. 005).[10] Test of significance of the two group mean composites was significant ($F = 20.99, p < .001$) suggesting good separation between criterion groups on the composites scores. The following lambda coefficients were obtained: Part A ($\lambda_1 = 3.5355$), Part B ($\lambda_2 = 3.6545$), Total ($\lambda_3 = -2.3610$), Duplication ($\lambda_4 = .4889$), Angulation ($\lambda_5 = .7204$), Time ($\lambda_6 = 1.3166$), Age ($\lambda_7 = -.0286$), and PIQ ($\lambda_8 = .2080$). The derived composite cut-off score was $Z \geqq 35.92$.

Table 9 presents the classification of *S*s on the basis of the new weightings (DF II). The table reveals that Discriminant Function II correctly classified 82% of the total sample ($N = 222$), with a false positive rate of 12% and a valid positive rate of 76%. By obtaining greater representation in both criterion groups, the high false positive rate on DF I (cross validation) decreased without appreciably altering the high valid positive rate in both studies. Further analysis again revealed that epileptic disorders accounted for the majority of the false negative errors.

Discriminant function II: Cross validation. The purpose of this investigation was to determine the predictive validity of DF II on a new sample of brain lesion and control *S*s. The study was carried out on 151 consecutive referrals to the laboratory during 1965, and consisted of 61 brain-lesion cases (Group I) and 90 psychiatric, general medical, and normal controls (Group II). Group I was represented equally by vascular, neoplastic, traumatic, and convulsive disorders. Group II was largely represented by psychiatric and general medical cases. Classification was made on the basis of the lambda coefficients derived from Discriminant Function II. The same composite cut-off score was again employed ($Z \geqq 35.92$). The results are presented in Table 10 and show that roughly 79% of the *S*s were correctly classified by this function (DF II). This second cross validation (DF II) showed considerably less shrinkage (3%) than the initial cross validation on DF I (14%). The results also showed a false positive rate of 16.67% and a valid

[9] Two *S*s who were originally diagnosed as epileptic were reclassified as nonorganic on the basis of additional information provided by the Neurology staff. The author extends his appreciation to Richard Weaver (Department of Neurology) and Lamar Roberts (Chief, Division of Neurosurgery) for their help in reevaluating some of the equivocal neurological cases.

[10] The computer analyses were run by the staff of the Computing Center, University of Florida.

TABLE 9

PREDICTIVE CLASSIFICATIONS BY USE OF DISCRIMINANT FUNCTION II (1963, 1964 SAMPLES) RESTANDARDIZATION ($N = 222$)

Interval composite scores	Nonorganics ($N = 117$)	Organics ($N = 105$)
76–79		1
72–75		4
68–71		6
64–67		9
60–63		3
56–59	1	4
52–55		7
48–51	2	14
44–47	3	12
40–43	2	11
36–39	6	9
32–35	8	6
28–31	33	9
24–27	34	7
20–23	25	3
16–19	3	

Note.—Composite cut-off score: $Z \geqq 35.92$; overall hits = 82%, valid positives = 76%, false positives = 12%.

TABLE 10

CROSS VALIDATION OF DISCRIMINANT FUNCTION II (1965 SAMPLE)

Interval composite scores	Nonorganics ($N = 90$)	Organics ($N = 61$)
80.92–85.91		1
75.92–80.91		1
70.92–75.91		4
65.92–70.91	1	8
60.92–65.91		2
55.92–60.91	1	2
50.92–55.91	1	2
45.92–50.91	1	5
40.92–45.91	6	12
35.92–40.91	5	7
30.92–35.91	24	4
25.92–30.91	30	6
20.92–25.91	17	6
15.92–20.91	4	1

Note.—Composite cut-off score: $Z \geqq 35.92$; overall hits = 79%, valid positives = 72%, false positives = 17%.

positive rate of 72.13%. With respect to the false positive errors, the present cross validation revealed a much smaller increase in errors (4%) than occurred with the original cross validation (29%). In summary, Discriminant Function II demonstrated adequate predictive validity on a new sample of brain-lesion control *Ss*, with only a slight increase in false positive and false negative rates of misclassification.

On the basis of these findings it was decided to combine the data of this cross-validated sample ($N = 151$) with the data on Discriminant Function II ($N = 222$), and to compute a new and final discriminant function based on even larger representative cases.

Discriminant function III: Restandardization (1966). Table 11 presents the means and differences between the organic ($N = 157$) and nonorganic ($N = 210$) criterion groups on each of the discriminant predictor variables for the combined samples (1963–1966).[11] Inspection of this table reveals significant group differences on each variable. Although the neurological group was both older ($t = 2.05$, $p < .05$) and less intelligent ($t = 2.15$, $p < .05$) their mean level of intelligence was within the Normal

TABLE 11

MEANS AND DIFFERENCES BETWEEN CRITERION GROUPS ON THE PREDICTOR VARIABLES RESTANDARDIZATION GROUP (DF III)

Variables	Nonorganic ($N = 210$)	Organic ($N = 157$)	t
Part A	3.67	11.76	9.56*
Part B	3.14	7.73	10.66*
Total	6.81	19.50	10.35*
Duplication	.53	3.11	4.48*
Angulation	.94	2.65	2.91*
Time	.64	1.43	2.23*
Age	37.00	41.66	2.05*
Performance IQ	101.20	94.65	2.15*

* All t values significant at $p \leqq .05$.

Range ($\bar{X} = 94.65$) and their mean age was not high ($\bar{X} = 41.66$).

The discriminant function analysis was

[11] Three *Ss* who were originally diagnosed as epileptic were reclassified as nonorganic on the basis of additional information provided by the Neurology staff. The Neurology Service also recommended that six additional patients from the organic group be dropped temporarily from this study for lack of definitive neurological evaluation. These *Ss* were placed in a brain tumor suspect group for further outpatient work-up.

TABLE 12
PREDICTIVE CLASSIFICATIONS BY USE OF DISCRIMINANT FUNCTION III (1963–1966) RESTANDARDIZATION
(N = 367)

Interval composite scores	Nonorganics (N = 210)	Organics (N = 157)
28.15–29.84		2
26.46–28.14		0
24.77–26.45		2
23.08–24.76		5
21.39–23.07		12
19.70–21.38		17
18.01–19.69	2	7
16.32–18.00	3	12
14.63–16.31	4	15
12.94–14.62	7	23
11.25–12.93	6	14
9.56–11.24	21	16
7.87–9.55	43	10
6.18–7.86	64	12
4.49–6.17	56	8
2.80–4.48	4	2

Note.—Composite cut-off score: $Z \geq 11.25$; overall hits = 81%, valid positives = 70%, false positives = 10%.

computed on an IBM 709 Computer with the use of the UCLA BIMED program (No. 005). Test of significance of the two group mean composites was significant (F = 26.52, $p < .001$), suggesting good separation between criterion groups on the composite Z scores. The following lambda coefficients were obtained:[12] Part A ($\lambda_1 = .9160$), Part B ($\lambda_2 = 1.3824$), Total ($\lambda_3 = -.6601$), Duplication ($\lambda_4 = .2991$), Angulation ($\lambda_5 = .3270$), Time ($\lambda_6 = .4415$), Age ($\lambda_7 = -.0289$), and PIQ ($\lambda_8 = .0500$). The derived composite cut-off score was $Z \geq 11.25$.[13]

Table 12 presents the classification of *S*s on the basis of the new weightings (DF III). Inspection of this table reveals that 81% of the total restandardization group (N = 367) were correctly classified by this new predictor function, with a false positive rate of 10% and a valid positive rate of 70%. In other words, 90% of the nonorganic controls and 70% of the organics were correctly classified by the restandardization function. These findings, in summary, satisfy the initial requirements concerning the predictive validity of this instrument.

Base-Rate Considerations

The purpose of the following analyses was to determine the utility or efficiency of the BRT (DF III) in settings in which extreme base-rate asymmetry might seriously affect the predictive validity of the instrument.

The first analysis was addressed to the problem of determining the probability of correct classification, giving a *positive* composite Z score on the BRT, under various theoretical base-rate populations. In more practical terms, the question asked was as follows: "How sure can you be on the basis of a positive test sign?" This problem involved the application of inverse probability given by Bayes' formula (Meehl & Rosen, 1955):

$$P_0 = \frac{P \cdot p_1}{P \cdot p_1 + Q \cdot p_2} \qquad (2)$$

P_0 = Probability that an individual is organic, given that his test score is positive.

P = Base rate of organics in the population.

Q = Base rate of nonorganics in the population.

p_1 = Proportion of organics identified by test ("valid positive" rate).

[12] The magnitude and direction of the lambda coefficients showed much variation between the separate discriminant function analyses (DF I–III). Most of the variation, however, occurred between DF I and DF II and was probably due to the lack of stability of the lambda estimates based on the smaller N in the standardization sample (DF I). When each of the coefficients (DF I–III) were converted to standard scores (separate analysis) it was found that the relative contributions of each variable remained essentially the same between DF II and DF III. The sample size for both of these standardization analyses was also appreciably higher.

[13] The three composite cut-off scores (DF I–III) varied primarily as a function of the way in which the lambda coefficients were converted from the computer program. The lambda coefficients for DF II and DF III were each multiplied by a constant of 1000 to avoid excessive use of decimals. The coefficients for DF I, however, were each multiplied by 1000 and then divided by the lowest lambda coefficient to avoid any values less than unity. This conversion accounted for the higher composite score obtained in this analysis (Z = 450.05).

TABLE 13

Number of Subjects Classified as Organic and Control by Discriminant Function III

Classification by DF III	Criterion classification: Organic N	Organic Percent	Combined control N	Combined control Percent	Total N classified by DF III
Organic	109	70	22	10	131
Combined control	48	30	188	90	236
Total in class	157	100	210	100	367

Probability of correct classification when $Z \geqq 11.25$ for base rates P and Q when $p_1 = .70$[a] and $p_2 = .10$[b].

Presumed base rates		P_0	Presumed base rates		P_0
$P = .10$	$Q = .90$	.44	$P = .60$	$Q = .40$	.91
$P = .20$	$Q = .80$	.64	$P = .70$	$Q = .30$	.94
$P = .30$	$Q = .70$	.75	$P = .80$	$Q = .20$	.97
$P = .40$	$Q = .60$	.82	$P = .90$	$Q = .10$	.98
$P = .50$	$Q = .50$	.88			

[a] p_1 = Valid positive rate, P = base rate of actual organics.
[b] p_2 = False positive rate, Q = base rate of actual nonorganics.

p_2 = Proportion of nonorganics misclassified by test ("false positive" rate).

The results of this analysis (Table 13) indicate that under conditions of extreme base-rate asymmetry in which the incidence of brain damage is low ($P = .10$, $Q = .90$), the diagnostician would be correct only 44 times in 100 when he predicted a brain lesion on the basis of a positive test sign with this instrument. This finding limits the usefulness of the BRT, with respect to this decision, for settings in which the incidence of brain injury is extremely low (e.g., mental hygiene clinics). In all other populations, however, the efficiency of this test would be high. In fact, its greatest usefulness would be in an inpatient medical setting ($P_0 = .91$) in which the incidence of brain disease is higher ($P = .60$). The probabilities of correct classification for various theoretical base-rate populations are plotted in Figure 3 and show how the likelihood of correct decisions increases markedly as the incidence of brain disease increases in the population ($P > Q$).

The second analysis was concerned with the percentage of correct classifications for positive and nonpositive BRT scores on projected samples ($N = 1000$) under various base-rate populations. This analysis provides a different framework for evaluating the potential usefulness of the BRT.

Table 14 presents the probable classification outcomes with this instrument for 1000 consecutive samples in a setting in which the psychologist typically functions ($P = .20$, $Q = .80$). By merely using the base rates in this setting, the diagnostician would be correct 80% of the time. His strategy would involve the nondifferential prediction of "nonorganic" in each case without being burdened with the administration, scoring, and interpretation of tests. The diagnos-

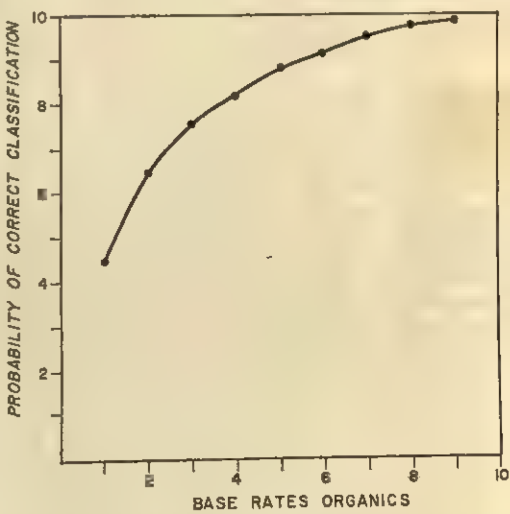

Fig. 3. Probability of correct classification when $Z \geqq 11.25$ (DF III) for combinations of base rates P and Q when $p_1 = .70$ and $p_2 = .10$.

TABLE 14

Probable Classification Outcome with Discriminant Function III for 1000 Projected Cases when $P = .20$ and $Q = .80$

Classification by function	Criterion classification				Total N classified by function
	Organic		Combined control		
	N	Percent	N	Percent	
Organic	140	70	80	10	220
Combined control	60	30	720	90	780
Total in class	200	100	800	100	1000

Note.—Valid positives = .70, false positives = .10. P = base rate of actual organics.

tician, however, who employed the BRT (DF III) systematically in this setting, would find his time worth the effort. When predicting "nonorganic" he would be correct 92% of the time (720/780). This represents a 12% increase in accuracy for the same type of decision (valid negative) when employing the BRT. Furthermore, the base rates alone would not detect a single brain lesion in this setting, whereas the test would result in the correct classification of 140 brain-injured *S*s ($p_1 = .64$).

Table 15 presents the probable classification outcomes ($N = 1000$) with this instrument in a typical medical inpatient setting similar to the present standardization population ($P = .35$, $Q = .65$). By using the base rates in this setting, the diagnostician would be correct 65% of the time by merely predicting "nonorganic" in each case. The BRT, however, would result in a 20% higher rate of classification for the same decision (585/690 = 85%). Furthermore, the test would correctly classify 80% of the brain disorders (245/310). Application of the base rates in this setting would not result in the identification of a single organic case! If the early detection of brain disease is considered essential to the medical treatment of a patient, then false negatives would represent more serious errors than false positives. In this case, the use of base rates would fail to detect a single brain lesion despite fair overall predictive validity ($H_T = 65\%$). On the other hand, the BRT would demonstrate better classification whether the psychologist predicted "organic" or "nonorganic," and would furthermore allow him to make some contribution in the setting in which he functioned.

TABLE 15

Probable Classification Outcome with Discriminant Function III for 1000 Projected Cases when $P = .35$ and $Q = .65$

Classification by function	Criterion classification				Total N classified by function
	Organic		Combined control		
	N	Percent	N	Percent	
Organic	245	70	65	10	310
Combined control	105	30	585	90	690
Total in class	350	100	650	100	1000

Note.—Valid positives = .70, false positives = .10. P = base rate of actual organics.

Cost Efficiency and Test Prediction

Assuming that differential error risks are involved between false negative and false positive decisions in organic brain assessment, then it would seem that the efficiency of any test should be evaluated further in terms of the relative costs of these two types of decision errors, and not merely against the prevailing base rates. The base rates assume implicitly that both types of decision errors are equally bad. In the preceding discussion, for example, it was felt that the failure to detect brain damage (a false nega-

tive error) was seemingly a more serious risk than to misclassify a nonorganic person (a false positive error), in that the failure to detect brain disorder could have serious implications with respect to the life of a human being, not to mention the additional medical expenses and repeated hospitalizations necessary if the disease process should remain undiagnosed. Furthermore, in view of the fact that the decision to employ the base rates would most likely result in the strategy to predict in favor of the higher Q values (i.e., when $P < Q$), then this comparison with the base rates might well lead to an incorrect decision regarding the acceptance or rejection of a given test. The reason is that the use of the base rates in this case would result in an absolute reduction of false positive errors at the expense of the more serious false negative errors.

Rimm (1963) has shown that some attempts to consider the relative error costs in dollars in test prediction can often reverse the decision regarding the rejection of a given test, despite unfavorable comparison with the base rates. The method he proposed included a utilization of the base rates (P and Q), the discriminatory efficiency of the test (p_1 and p_2), *and* a relatively simple formula involving these differential error costs.[14] This formula was defined as follows:

$$\text{Cost efficiency} = \frac{p_1 - R(1 - P)p_2}{P} \quad (3)$$

in which R represents the ratio of the cost of a false positive error to the cost of a false negative error.

In order to derive a numerical translation of these relative error costs, it was assumed that a false negative error would result in the long run in *twice* the investment of money and professional man hours as compared with a false positive error. Therefore the dollar cost ratio of a false positive to a false negative error was set at 1:2; that is, R was equal to ½ or .5.

To reexamine the BRT (DF III), given a valid positive rate of $p_1 = .70$ and a false positive rate of $p_2 = .10$, for the more typical clinical setting in which the incidence of brain disorder is $P = .20$, the cost of efficiency of this predictor function would be:

$$.70 - \frac{.50(.80).10}{.20} = .50$$

This value means that for every dollar that would have been spent paying for the base rate errors resulting from the prediction of "nonorganic" in every case, .50 dollar would have been *saved* as a result of employing the BRT (DF III). Or, in other words, for every dollar spent as a result of errors obtained by using the base rates, $1 - .50 = .50$ dollar would have been *spent* if the BRT had been used instead. Equally striking are the results for populations in which the incidence of brain disease was extremely low ($P = .10$). These findings are reported in Figure 4. The cost efficiency of the BRT, when $P = .10$, was .25, which indicates that even under extreme base rate asymmetry it would have been more efficient to employ the BRT. Figure 4 also shows that the superiority of the BRT increases as the incidence of brain disease increases in the population ($P > Q$).

Composition and Classification of the Criterion Groups

Table 16 presents the major subgroups within the organic and nonorganic criterion groups and the classification frequencies of each subgroup based on the discriminant function composite scores (DF III). The cutting line derived from analysis of DF III is presented to show the differential classification rates between subgroups. The nonorganic group was largely composed of psychiatric ($N = 120$) and medical disorders ($N = 51$) and a smaller group of normal *S*s ($N = 39$). Roughly 35% of the psychiatric subgroup was composed of schizophrenics. The highest rate of correct classification was obtained with the normal (97%) and psychiatric groups (90%). It was the medical subgroup that showed the greatest amount of misclassification (18%). Headaches, nausea, vomiting, and dizziness were the most characteristic presenting symptoms

[14] The reader is referred to this interesting article for a more detailed explanation of the problem and the mathematical steps involved (Rimm, 1963).

TABLE 16

PREDICTIVE CLASSIFICATIONS FOR SUBGROUPS WITHIN THE ORGANIC AND NONORGANIC CRITERION GROUPS BY USE OF DISCRIMINANT FUNCTION III

Interval composite scores	Organic (N = 157)							Nonorganic (N = 210)		
	SX	Neop	Vasc	Traum	ASCVD	Drug	Other	Norm	Med	Psych
28.15–29.84						2				
26.46–28.14										
24.77–26.45	1		1							
23.08–24.76		1	2		2					
21.39–23.07	1	3	3	1	3		1			
19.70–21.38	1	1	4	3	7	1				
18.01–19.69		2	2	2	1				1	1
16.32–18.00	3	1			6	1	1		1	2
14.63–16.31	1	2	4	3	4		1	1		3
12.94–14.62	6	1	4	4	4	2	2		4	3
11.25–12.93	3		5	3	3				3	3
9.56–11.24	4	4	3		3		2	3	5	13
7.87–9.55	1	1		2	3	3		7	11	25
6.18–7.86	6	2	1	2			1	13	15	36
4.49–6.17	4	1	1	1	1			15	10	31
2.80–4.48	1					1			1	3
Total	32	19	30	21	37	10	8	39	51	120
Hits	16	11	25	16	30	6	5	38	2	108
Misses	16	8	5	5	7	4	3	1	9	12

Note.—SX = seizures, Neop = neoplasms, Vasc = vascular, Traum = traumatic, ASCVD = arterio-sclerotic-vascular, Norm = normals, Med = medical, Psych = psychiatric.

in this group, although neurological examination failed to document the presence of brain disease in these patients.

The organic group was composed largely of vascular (N = 30), neoplastic (N = 19), traumatic (N = 21), arteriosclerotic (N = 37), and convulsive disorders (N = 32). Inspection of this table shows that the highest rate of correct classification occurred with the vascular (83%) and arteriosclerotic disorders (81%). On the other hand, the largest percentage of misclassification (false negatives) occurred with the convulsive disorders (50%). In order to determine more clearly the effects of disease classification on test performance, an analysis of variance was then computed on the discriminant composite scores for each subgroup (Hypothesis Eq. 3).

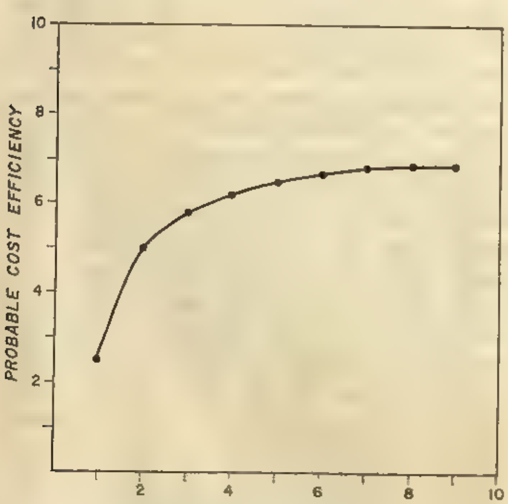

FIG. 4. Cost efficiency probabilities of DF III for combinations of base rates P and Q when p_1 = .70 and p_2 = .10.

Analysis of Classification of Brain Disease

The mean discriminant composite scores for the different lesion groups are presented in Table 17. Although each of the composite means was above the discriminant cutting line (Z = 11.25), the mean of the convulsive group fell only slightly above this value (Z = 11.58). An analysis of variance (Table 17) on the composite means revealed an overall separation between groups (F = 2.51, $p < .02$). Additional tests between

TABLE 17
DISCRIMINANT FUNCTION (III) COMPOSITE MEANS BY CLASSIFICATION OF BRAIN DISORDER

Type	SX	Neop	Vasc	Traum	ASCVD	Drug	Other
Composite mean	11.58	14.74	16.09	14.07	15.97	15.28	13.76
N	32	19	30	21	37	10	8

Analysis of variance of composite predictor scores

Source	SS	df	MS	F	p
Between	442.89	6	73.82	2.51	<.02
Within	4416.21	150	29.44		
Total	4859.10	156			

individual group means, however, showed that these differences were largely due to the effects of the convulsive group. These findings lend only partial support for the hypothesis that classification of lesion is a meaningful independent variable within the concept of "brain damage."

Analysis of Generalized Effects

This analysis was essentially an extension of the earlier analysis addressed to the problem of whether the composite scores were independent of area of involvement. In this analysis *S*s were classified only according to laterality of lesion, with 34 left hemisphere and 34 right hemisphere cases. The remaining group of indeterminate lesions ($N = 89$) consisted largely of bilateral and diffuse brain-lesion cases. These data are presented in Table 18. Disease classification was controlled for between left and right hemispheric cases due to the effects of this variable on test performance. In other words, both hemispheric groups had the same frequency of vascular, neoplastic, traumatic, arteriosclerotic, and convulsive disorders. Inspection of Table 18 shows that the discriminant composite means were all well above the cutting line ($Z \geqq 11.25$). Analysis of variance of the composite means, however, showed no overall differentiation between area of involvement ($F < 1$). In fact the composite mean for the left hemisphere group ($Z = 14.62$) was almost the same as for the right hemisphere group ($Z = 14.60$). This finding lends additional support for Hypothesis 2.

TABLE 18
DISCRIMINANT FUNCTION (III) COMPOSITE MEANS BY AREA OF BRAIN INVOLVEMENT

Area of damage	N	Composite mean
Left hemisphere	34	14.62
Right hemisphere	34	14.60
Indeterminate hemisphere	89	14.48

Analysis of variance of composite predictor scores

Source	ss	df	MS	F	p
Between	.61	2	.31	.009	ns
Within	4858.49	154	31.55		
Total	4859.10	156			

DISCUSSION

The present findings lend considerable support for the predictive validity of the BRT. Although classification shrinkage occurred after the cross validation of DF I, the subsequent cross validation and restandardization analyses revealed a more stable level of predictive classification. This was probably due to increased representation of cases within both criterion groups which helped to stabilize the lambdas for each of the predictor variables. More impressive, however, is the fact that the initial cross validation (DF I) showed only a 14% shrinkage in overall hits despite the fact that a new population was sampled, different examiners were used, and further, that the original standardization (DF I) was based on a relatively small sample of cases ($N = 122$). The initial cross validation also revealed that the false positive errors accounted for most of this shrinkage, varying

from 5% on the original standardization (DF I) to 29% after cross validation (DF I). On the subsequent validations, however, the false positive rate never went beyond 17%, and finally stabilized around 10% on the final restandardization (DF III).

The false negative rate, on the other hand, was more consistent between each of the validation analyses, varying from 19% on the initial standardization (DF I) to 30% on the final restandardization (DF III). These errors, however, accounted for the majority of misclassifications throughout the study.

Further analysis of the false negative misclassifications, on the entire brain-injured standardization group (DF III), demonstrated clearly that the *classification* of brain disorder was a significant independent variable within the neurological group. It is evident from Tables 16 and 17 that convulsive disorders accounted for the majority of false negative errors. In fact, a separate analysis of hits within the seizure group, according to age, revealed that 85% of the epileptics below age 35 were missed by the BRT. It was only in the higher age ranges that the BRT correctly classified the majority of seizure disorders. These latter cases were largely selected from the initial Veterans Administration standardization population (DF I). This finding suggests that age, that is, length of seizures or chronicity, might lead to more generalized brain dysfunction in man. With respect to the high incidence of false negatives within the convulsive group it should also be pointed out that this disorder seldom results in any demonstrable structural change in the brain. The neurological studies on these patients in the present study (e.g., arteriography, pneumography, brain scans, etc.) were uniformly negative. Their only positive neurological findings were EEG abnormalities and a history of fits. On the basis of these findings, one could argue against the inclusion of convulsive disorders in the neurological group, particularly at ages under 35. This procedure would, in turn, increase the percentage of valid positives, and restrict the purpose of the test to organic disorders involving structural changes in the brain. This is essentially the procedure that neurology and neurosurgery follow in evaluating convulsive disorders, that is, to determine whether the seizures are secondary to neoplastic, vascular, or traumatic disease.

Similar problems were encountered in the analysis of false negative misclassifications within the neoplastic group. Although the majority of neoplastic lesions were correctly classified, five of the eight misclassifications in this group occurred with lesions not involving the cerebral cortex (i.e., pituitary and cerebellar tumors), and one error was associated with a small meningioma. In other words, only two neoplastic lesions, involving the cerebral cortex, were misclassified with the BRT (DF III). Future research might likewise dictate that this instrument be restricted to lesions involving the cerebral cortical structures. Similar diagnostic procedures are followed in medicine. The majority of laboratory tests in neurology, for example, are designed for specific types and/or areas of brain pathology (Merritt, 1959). Furthermore, it is with cortical lesions that many of these laboratory tests have shown an increase in false negative classification (Brosin, 1959).

The highest percentage of valid positives occurred within the vascular and arteriosclerotic-degenerative (ASCVD) brain-lesion groups (81%). This was probably due to the infrequency of focal disease in these patients. The majority of *S*s within the vascular group had widespread involvement associated with one hemisphere whereas many of the *S*s within the ASCVD group had diffuse cerebral involvement.

The analysis on *types* of brain disorder provided further support for the importance of isolating other variables within the concept of "brain disorder." Acute brain lesions showed greater impairment on their composite Z scores (DF I) than did the relatively static brain-lesion group. Although the difference between the acute and chronic-static groups was not significant, the trend was in the direction predicted. The inclusion of prefrontal cases within the chronic-static group in the present study might have contributed to the suppression of differences between these two groups. In the Fitzhugh et al. study (1961), chronic epileptics comprised the majority of patients

within the chronic-static group. Apparently the effects of prefrontal damage were greater than the effects of longstanding epilepsy. The fact that the majority of classification errors occurred within the relatively static group (DF I) is important with respect to the selection of neurological patients in the standardization of a predictive test of brain dysfunction. The present findings showed that test performance varied as a function of both type and classification of brain disorder. Failure to control for the effects of these variables could therefore bias the outcome of any predictive validation study.

Area of brain involvement, on the other hand, was not shown to be related to test performance. The mean predictor composites (Tables 6 and 18) were well above the cutting line in both analyses, and they failed to reveal any overall difference whether damage occurred in the frontal regions, left temporoparietal regions or right temporoparietal regions. These findings were consistent with previous results which have shown nonspecific effects on *complex* visuo-perceptual tests (Teuber, 1959; Teuber & Weinstein, 1956). Teuber (1959) hypothesized that performance on these complex perceptual measures depends on a number of different psychologic functions, each of which is crucial. Hence, any lesion sufficient to affect one of these functions may lead to a significant *general* impairment. Three postulated levels of performance on the BRT were separately analyzed in an earlier study (Satz, 1966). Level I was defined as the perception of the stimulus design as presented by *E*. Level II was defined as perception of the rotated stimulus image; and Level III was defined as motoric translation of the rotated perceptual image. These three levels apparently interact in that an impairment at Level I, that is, an inability to correctly perceive the stimulus design, leads to a breakdown in performance at the two higher levels. In like manner, an error at Levels II or III does not, however, necessitate an error at Level I; for the *S* may be able to correctly perceive the initial stimulus design, but may not be able to perform correctly at the two more difficult levels (Levels II and III). Evidence for this test behavior was found in the present study in which several of the organic *S*s failed by merely reproducing *E*'s stimulus designs (Duplication error). A reproduction of the stimulus design, however, indicates, by definition, correct perception at Level I. This type of error performance parallels rather closely the concept of "stimulus boundedness" cited by other investigators (Goldstein, 1959; Hemmendinger, 1953). It also suggests that different levels of perception might be involved in performance in a task such as the BRT; first, a lower level recognition-discrimination system; and second, a higher level system involving more complex integrative perceptual processes. Bortner and Birch (1960, 1962) have recently advanced support for a similar hypothesis of levels in perception. The preceding analysis of the BRT suggests that an error at any of the three levels will lead to general impairment on any test item. If it could be demonstrated that different brain lesions have different effects on each of these levels, then one could account for the nonspecific effects obtained with this test. A lesion sufficient to upset one of the postulated levels would lead to significant *general* impairment.

An additional part of this study was addressed to the utility or efficiency of the BRT within various theoretical base rate populations. The results showed rather clearly the effects of extreme base rate asymmetry ($P \ll Q$) on this instrument. In settings in which the incidence of brain disease was low ($P \leq .20$), the decision to employ the BRT would depend on a number of factors in addition to overall hits, percentage of valid positives and percentage of false positives. One of the factors involved the varying incidence of brain disease in different clinical settings (i.e., base rates). If, for example, the incidence of brain disease was low in a particular setting ($P = .20$, $Q = .80$), the diagnostician could be assured of an overall hit rate of 80% by merely using the higher base rate value (i.e., Q) and predicting "nonorganic" in each case. This procedure would require no test administration, scoring, or interpretive time. Furthermore, the percentage of overall hits would appear to be approximately the same whether the test (DF III) or base rates were

employed (81% versus 80%). If the incidence of brain disease was even lower (i.e., $P = .10$), the advantages of employing the base rates would appear even more striking (81% versus 90%). However, the percentage of overall hits (H_T) is a value that is often misleading in test prediction and which can be spuriously inflated by a much larger N in the criterion group with the higher valid positive or valid negative rate. That is why it is essential to have information on the percentage of valid positives and percentage of false positives for any predictive test. Furthermore, in comparing the relative efficiency of a test against the prevailing base rates, it is additionally important to determine the predictive outcomes on projected samples. This approach was followed in the present study (Tables 14 and 15). For the setting in which the incidence of brain disease was low ($P = .20$, $Q = .80$), it was shown that it would be risky to conclude that the base rates were superior to the BRT (DF III) on the basis of overall hits (Table 14). When the test was examined on 1,000 projected cases, given $P = .20$, it was discovered that the diagnostician who employed the test systematically in this setting would find his time well spent. When predicting "nonorganic" he would be correct 92% of the time. This outcome represented a 12% increase in accuracy over the base rates for the same decision statement. Further, the base rates would not detect a single brain lesion in this setting, whereas the test would correctly identify 140 brain-injured *S*s. Similar, although less striking, advantages of the test were found in the more extreme base rate populations ($P = .10$, $Q = .90$). In other words, the original comparison between the BRT and base rates, on overall hits (81% versus 80%), was shown to be misleading without additional analysis on projected samples involving both percentage of valid positives and percentage of false positives.

A second approach, for determining the relative efficiency of the test, was addressed to the more practical question: How sure could you be on the basis of a positive test sign with this instrument under varying base rate populations? This problem has been discussed by Meehl and Rosen (1955) and involved the application of inverse probability. The results of this analysis (Figure 3) showed that only in settings in which the incidence of brain disease was extremely low ($P = .10$, $Q = .90$) would the diagnostician be wrong more often than right in predicting the presence of brain disorder, given a positive score on the BRT ($Z \geqq 11.25$). In all other clinical settings the diagnostician could be quite confident of his prediction of brain dysfunction on the basis of a positive composite score with this instrument. In other words, the risk of a false positive error was shown to be more likely in settings similar to mental health clinics in which the incidence of brain disease is low. In other settings, particularly hospitals and inpatient medical services, the probability of being correct on the basis of a positive BRT score would be high.

A third approach, for determining the utility or efficiency of the BRT, was addressed to the differential risks between false positive and false negative decision errors. For example, in settings in which the incidence of brain disease was low ($P < Q$), the strategy to employ the higher base rate value (i.e., Q) would have resulted in an absolute reduction of false positive errors at the expense of the false negative errors. In other words, without the test, not a single brain-lesion case would have been detected. This decision error (false negative), however, was felt to represent a more serious risk than the misclassification of a normal person (false positive) because the failure to detect the presence of brain disease could lead to grave consequences and additional cost for an individual should the disease process remain undiagnosed. Implicit to the base rates is the assumption that both decision errors are equally risky. Rimm (1963) has shown that attempts to consider the relative error costs in dollars for predictive tests can often reverse the decision regarding the rejection of a given test, despite unfavorable comparison with the base rates. In the present study the attempt to translate the ratio of these differential error risks into numerical valuation provided the third framework for evaluating the efficiency of the BRT. These cost efficiency analyses provided additional support for the superiority of the

BRT, even for settings in which the incidence of brain disease was extremely low ($P = .10$, $Q = .90$).

The final interpretations with respect to the usefulness of this predictive instrument must perforce be left to the decision of the particular diagnostician. The preceding analyses on predictive efficiency merely represented different frameworks for asking a number of different questions about the possible usefulness of this test. Although the predictive validity of the final restandardization was high (DF III), it was shown that the efficiency of the BRT was somewhat limited by the effects of extreme base rate asymmetry ($P \ll Q$), but only when the psychologist was concerned about the probability of being correct on the basis of a positive test sign. This finding should be of some concern to psychologists working in community mental health centers. On the other hand, if it was decided that in this setting the risk of false negative error was more costly than a false positive error, then the strategy to employ this test could be justified. The diagnostician could at least be confident that his false positive rate would be small (i.e., 10%) and, should the more serious problem of brain pathology be present, he would, in the majority of cases, detect it. But again, these decisions would have to be weighed against the additional time and expense involved in test administration and scoring.

Before concluding it should be emphasized that the predictive validity of this multivariate instrument is by no means fully demonstrated. Additional cross-validational studies are still needed, preferably in other settings. More serious is the problem of defining clearly the predictive purposes of the test, particularly with respect to classification and area of brain involvement. Should the criterion definitions pertain only to those organic brain lesions which cause structural damage restricted to the cerebral cortex? This criterion specification would obviously exclude the majority of convulsive disorders and neoplasms situated in lower centers of the brain. More research is needed before these problems can be resolved.

A final word should be given to objections which might arise on the utility of an instrument that purports to measure only the presence or absence of brain disorder. The point is that before the psychologist can begin to make second-level inferences concerning the laterality or localization of brain lesions, he must first demonstrate a strong likelihood for the presence of brain dysfunction. Only then can he proceed, using additional tests, to other levels of decision inference. The decision process in neurology, on the other hand, is not always restricted to a step-wise level of inferences. There are test procedures (e.g., angiography) in which visualization of the lesion leads, by definition, to detection and often provides additional information concerning the type of lesion and whether or not it is localized. The present investigation was carried out to examine the complexities of the initial stages of this decision process for the psychologist.

REFERENCES

Aita, J. A., Armitage, S. G., Reitan, R. M., & Rabinowitz, A. The use of certain psychological tests in the evaluation of brain injury. *Journal on General Psychology*, 1947, 37, 25–44.

Armitage, S. G. An analysis of certain psychological tests used in evaluation of brain-injury. *Psychological Monographs*, 1946, 60 (1, Whole No. 277).

Arrigoni, G., & De Renzi, E. Constructional apraxia and hemispheric locus of lesion. *Cortex*, 1964, 1, 170–197.

Bender, M. B., & Teuber, H. L. Spatial organization of visual perception following injury to the brain. *Archives of Neurology and Psychiatry*, 1948, 59, 39–62.

Bortner, M.,& Birch, H. G. Perceptual and perceptual-motor dissociation in brain-damaged patients. *Journal of Nervous and Mental Diseases*, 1960, 130, 49–53.

Bortner, M., & Birch, H. G. Perceptual and perceptual-motor dissociation in cerebral palsied children. *Journal of Nervous and Mental Diseases*, 1962, 134, 103–108.

Brosin, H. W. Psychiatric conditions following head injury. In S. Arieti (Ed.), *American handbook of psychiatry*. Vol. II. New York: Basic Books, 1959. Pp. 1175–1203.

Chorost, S. G., Spivak, G., & Levine, M. Bender-Gestalt rotations and EEG abnormalities

in Children. *Journal of Consulting Psychology*, 1959, 23, 559–560.

COSTA, L. D., & VAUGHAN, H. G. Performance of patients with lateralized cerebral lesions. I: Verbal and perceptual tests. *Journal of Nervous and Mental Diseases*, 1962, 134, 162–168.

DENNERLL, R. D. Cognitive deficits and lateral brain dysfunction in temporal lobe epilepsy. *Epilepsia*, 1964, 5, 177–191.

ESTES, B. W. A note on the Satz-Mogel Abbreviation of the WAIS. *Journal of Clinical Psychology*, 1963, 19, 101.

FISHER, R. A. The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, 1936, 7, 179–188.

FITZHUGH, K. B., FITZHUGH, L. C., & REITAN, R. M. Psychological deficits in relation to acuteness of brain dysfunction. *Journal of Consul ing Psychology*, 1961, 25, 61–66.

GARRETT, H. E. The discriminant function and its use in psychology. *Psychometrika*, 1943, 8, 65–79.

GOLDSTEIN, K., & SCHEERER, M. Abstract and concrete behavior; An experimental study with special tests. *Psychological Monographs*, 1941, 53, (2, Whole No. 239).

GOLDSTEIN, K. Functional disturbances in brain damage. In S. Arieti (Ed.), *American handbook of psychiatry*. Vol. 1. New York: Basic Books, 1959. Pp. 770–797.

GOULDEN, C. H. *Methods of statistical analysis*. New York: Wiley, 1956.

GRAHAM, F. K., & KENDALL, B. S. Memory-for-Designs Test: Revised general manual. *Perceptual Motor Skills*, 1960, 11, 147–188.

GRIFFITH, R. M., & TAYLOR, V. H. Incidence of Bender-Gestalt figure rotations. *Journal of Consulting Psychology*, 1960, 24, 189–191. (a)

GRIFFITH, R. M., & TAYLOR, V. H. The effect of mental illness on intelligence test scores. *Journal of Clinical Psychology*, 1960, 16, 352. (b)

HALSTEAD, W. C. *Brain and intelligence*. Chicago: University of Chicago, 1947.

HEBB, D. O. *The organization of behavior*. New York: Wiley, 1949.

HEILBRUN, A. B. Psychological test performance as a function of lateral localization of cerebral lesion. *Journal of Comparative and Physiological Psychology*, 1956, 49, 10–14.

HEMMENDINGER, L. Perceptual organization and development as reflected in the structure of Rorschach test responses. *Journal of Projective Techniques*, 1953, 17, 162–170.

HUNT, H. F. A practical clinical test for organic brain damage. *Journal of Applied Psychology*, 1943, 27, 375–386.

JASPER, H., KERSHMAN, J., & ELVIDGE, A. Electroencephalograph in head injury. *Research Publication Association, Research Nervous and Mental Disease*, 1945, 24, 383–420.

LASHLEY, K. S. In search of the engram. In F. A. Beach et al., *The neuropsychology of Lashley*. New York: McGraw-Hill, 1960. P. 494.

MERRITT, H. H. *Textbook of neurology*. Philadelphia: Lea & Febiger, 1959.

MEEHL, P. E., & ROSEN, A. Antecedent probability and the efficiency of signs, patterns, or cutting scores. *Psychological Bulletin*, 1955, 52, 194–216.

MEYER, V. Psychological effects of brain damage. In H. J. Eysenck (Ed.), *Handbook of abnormal psychology*. New York: Basic Books, 1961. Pp. 529–566.

MILNER, B. Psychological functions of temporal lobe. *Psychological Bulletin*, 1954, 51, 42–62.

MOGEL, S., & SATZ, P. An abbreviation of the WAIS for clinical use: An attempt at validation. *Journal of Clinical Psychology*, 1963, 19, 298–300.

PASCAL, G. R., & SUTTELL, B. J. *The Bender-Gestalt Test*. New York: Grune & Stratton, 1951.

PAUKER, J. D. A split-half abbreviation of the WAIS. *Journal of Clinical Psychology*, 1963, 19, 98–100.

PENFIELD, W., & ROBERTS, L. *Speech and brain mechanisms*. Princeton: Princeton University, 1959.

REITAN, R. M. Certain differential effects of left and right cerebral lesions in human adults. *Journal of Comparative and Physiological Psychology*, 1955, 48, 474–477.

REITAN, R. M. Investigation of the relationships between "psychometric" and "biological" intelligence. *Journal of Nervous and Mental Diseases*, 1956, 123, 536–541.

REITAN, R. M. Validity of the Trail Making Test as an indicator of organic brain damage. *Perceptual and Motor Skills*, 1958, 8, 271–276.

REITAN, R. M. Psychological deficit. *Annual Review of Psychology*, 1962, 13, 415–444.

RIMM, D. Cost efficiency and test prediction. *Journal of Consulting Psychology*, 1963, 27, 89–91.

SATZ, P. *A block rotation task and multivariate statistical procedure for the diagnosis of organic brain disorder*. Unpublished doctoral dissertation, University of Kentucky, 1963.

SATZ, P. Specific and nonspecific effects of brain lesions in man. *Journal of Abnormal Psychology*, 1966, 71, 65–70.

SATZ, P., & MOGEL, S. An abbreviation of the WAIS for clinical use. *Journal of Clinical Psychology*, 1962, 18, 77–80.

SEMMES, J., WEINSTEIN, S., GHENT, L., & TEUBER, H. L. *Somatosensory changes after penetrating brain wounds in man*. Cambridge: Harvard University Press, 1960.

SHAPIRO, M. B. Experimental studies of a perceptual anomaly. I. Initial experiments. *Journal of Mental Science*, 1951, 97, 90–110.

SHAPIRO, M. B. Experimental studies of a perceptual anomaly. II. Confirmation and explanatory experiments. *Journal of Mental Science*, 1952, 98, 605–617.

SHAPIRO, M. B. Experimental studies of a perceptual anomaly. III. The testing of an explanatory theory. *Journal of Mental Science*, 1953, 99, 394–409.

STRAUSS, A. A., & LEHTINEN, L. E. *Psychopathology and education of the brain injured child*. New York: Grune & Stratton, 1950.

Teuber, H. L. Effects of brain wounds implicating right or left hemisphere in man. In V. B. Mountcastle (Ed.), *Interhemispheric relations andcerebral dominance*. Baltimore: Johns Hopkins Press, 1962.

Teuber, H. L., & Weinstein, S. Ability to discover hidden figures after cerebral lesions. *A.M.A. Archives of Neurology and Psychiatry*, 1956, **76**, 369–379.

Teuber, H. L. Some alterations in behavior after cerebral lesions in man. In A. D. Bass (Ed.), *Evolution of nervous control*. Washington, D. C.: American Association for the Advancement of Science, 1959. Pp. 157–194.

Watson, C. G. Evidence on the utilities of three WAIS short-forms. *Journal of Consulting Psychology*, 1966, **30**, 181.

Wechsler, D. *The measurement and appraisal of adult intelligence*. Baltimore: Williams & Wilkins, 1958.

Weinstein, S., & Teuber, H. L. Effects of penetrating brain injury on intelligence test scores. *Science*, 1957, **125**, 1036–1037.

Williams, H. L., Lubin, A., Gieseking, C., & Rubinstein, I. The relation of brain injury and visual perception to block design rotation. *Journal of Consulting Psychology*, 1956, **20**, 275–280.

Williams, H. L., Gieseking, C., & Lubin, A. Interaction of brain injury with peripheral vision and set. *Journal of Consulting Psychology*, 1961, **25**, 543–548.

Yates, A. J. The validity of some psychological tests of brain damage. *Psychological Bulletin*, 1954, **51**, 359–379. (a)

Yates, A. J. *An experimental study of the block design rotation effect with special reference to brain damage*. Unpublished doctoral dissertation, University of London, 1954. (b)

(Received April 26, 1966)

Vol. 80, No. 22 Whole No. 630, 1966

Psychological Monographs: General and Applied

SUBSTANTIVE DIMENSIONS OF SELF-REPORT IN THE MMPI ITEM POOL[1]

JERRY S. WIGGINS

University of Illinois

Starting with the original item-content classifications of Hathaway and McKinley, both psychometric and intuitive procedures were employed in the development of a set of scales designed to be internally consistent, moderately independent, and representative of the major substantive clusters that appeared to exist in the total MMPI item pool. Although not constructed by the strategy of contrasted groups, the 13 MMPI content scales nevertheless exhibited significant variability of mean scale scores when diverse normal, college, and psychiatric populations were compared. Moreover, when multivariate analytic procedures were employed in the diagnostic classification of psychiatric inpatients, the content scales were found to be as promising as the conventional clinical scales which were derived for this specific purpose. The factorial structure of the content scales was related to that of the clinical scales and found to lend support to the interpretation of the first 2 factors of the MMPI as "ego resiliency" and "control." An example was provided of the manner in which content scales may be employed as a supplement to interpretation of the MMPI clinical scales.

THE concept of item *content* has enjoyed neither precise specification nor active empirical exploration in the recent history of objective personality assessment. This situation may, in part, be attributed to the general disrepute into which "face validity" has fallen as a validity criterion (APA, 1954; Stagner, 1958) and to the tendency to regard responses to ambiguous items as reflecting "dynamic" aspects of personality. Meehl's (1945) now classic empirical manifesto raised the hope that dynamic aspects of personality might be assessed by means of true-false item pools superficially bearing little resemblance to the criterion at hand. A not infrequently encountered corollary of this belief is the superstition that knowledge of the content of an empirical scale may somehow vitiate the mysterious mediating process that links scale scores to empirical criteria.

Jackson and Messick's (1958) influential distinction between content and style in personality assessment was partly motivated by their desire to measure the former class of variables with more precision. Their methodological innovations enabled them to separate, within limits, components of content and style in the MMPI (Jackson & Messick, 1961). Several studies later, their view of content has a wistful tone: "Actually, we are very much concerned with measuring content, but content—like a tarpon being hunted by a spear fisherman at ten fathoms—usually appears somewhat closer, larger, and more easily captured than is actually the case [Jackson & Messick, 1962]." Without denying the rather poor showing that content made in their studies, it should be noted that their criterion for content demanded interitem consistencies within the

[1] The early phases of this research were supported in part by a research grant, MH 07042-01, from the National Institute of Mental Health. The final phases of the project were supported by a grant from the University Research Board of the Graduate College, University of Illinois.

The writer is indebted to Lewis R. Goldberg and Nancy Wiggins for their helpful criticisms of an earlier version of this monograph.

MMPI clinical scales that survived the partialing out of two potent sources of content-confounded stylistic variance—"acquiescence" and "social desirability." As will be indicated later, the assessment strategy of contrasted criterion groups (Wiggins, 1962) which was employed in the development of the MMPI clinical scales cannot be expected to insure content homogeneity within empirical scales.

The most nihilistic position with respect to personality test item content has been taken by Irwin Berg, the originator of the Deviation Hypothesis (Berg, 1955, 1959, 1961). The assessment *strategy* of statistical differentiation between deviant and normative groups has impressed Berg as being so fundamental to personality measurement as to render unimportant the item *content* whereby this differentiation is achieved (Berg, 1959). In many ways this position is a restatement of the pragmaticism of the empirical movement (Meehl, 1945) with a non sequitur corollary which makes the blindness of blind empiricism a virtue. Berg's main point with respect to content seems to be that any given content may be considered *in principle* to be as effective for a predictive task as any other content and that recourse should be made to empirical evidence as the final arbiter. He is careful to note that this does not mean that "... any item is just as good as every other for discriminative purposes [Berg, 1959, p. 89]." However, he tends to overstate his case:

> ...one should be able to construct the MMPI scales from the Strong Interest Blank and the Strong occupational scales from the MMPI items by using the same technique. Or, for that matter, one should be able to develop the scales of both tests from almost any hodge-podge of a similar number of items.... Given enough deviant responses and clean criterion groups, one should be able to duplicate any existing personality, interest, occupational and similar scales without regard to particular item content [Berg, 1955, p. 70].

The basis of the above inference is not clear since Berg is unable to provide even a rudimentary rationale whereby one might be able to predict the suitability of a given content for a given assessment. Similarly, when he states: "... the carefully described 26 categories of test item content employed by the MMPI are probably irrelevant for clinical measurement purposes [Berg, 1961, p. 361]," it is not clear how one might know this in advance of empirical test. One may argue that, *in principle*, alternative and equally effective item pools might be discovered for any given prediction situation and such a principle cannot be disproved. To prejudge a given item pool requires a theory of content, however, and Berg's contribution to this enterprise has been mainly a negative one (Norman, 1963).

In light of the foregoing discussion it is not surprising that the 26 content categories involved in the original classification of the MMPI item pool have received little attention in the literature (Wiggins & Vollmar, 1959). The test authors themselves (Hathaway & McKinley, 1940) were reluctant to attribute much significance to either selection or classification of item content, although their aim "... that more varied subject matter be included to obtain a wider sampling of behavior of significance to the psychiatrist [p. 249]" seems clearly to have been met. The content categories, themselves, have not excited the curiosity of many of the authors who have contributed to the nearly 1,000 articles (Dahlstrom & Welsh, 1960) on the MMPI that have since appeared.

While the academic and professional community have seen fit to ignore or denigrate the content of the MMPI, other segments of our society ("subjects") have been less quiescent (see *American Psychologist*, APA, 1965). Viewed from the other side of the desk, the 566 items of the MMPI appear to represent a massive invasion of privacy. Appeals to the principles of empiricism (Gordon, 1965) serve only to emphasize the insensitivity of the professionals involved, and such appeals hardly justify the use of any particular set of items for a given selection purpose. Attempts to placate the public by removing the more "offensive" items from the MMPI pool (Braaten, 1965) cannot be justified on a scientific basis. In short, a legitimate issue has been raised concerning the content

of personality inventory items, and the scientific and professional community has been stirred from an undeserved complacency. The viewpoint that a personality test protocol represents a communication between the subject and the tester (or the institution he represents) has much to commend it; not the least of which is the likelihood that this is the frame of reference adopted by the subject, himself.

The present study represents a first step in the direction of clarifying the content of the MMPI item pool. Starting with the original content classifications of Hathaway and McKinley, both psychometric and intuitive procedures were employed in the development of a set of scales designed to be internally consistent, moderately independent, and representative of the major substantive clusters that appeared to exist in the total MMPI item pool.

There have been a number of previous attempts to provide bases for regrouping MMPI items in ways other than that provided by the standard empirical scales. For the most part, these studies have used existing empirical scales as the basis for further regrouping. Homogeneous subgroupings of items within each of the standard empirical scales have been identified on a rational basis by Harris and Lingoes (1955) and on a factor analytic basis by Comrey (1957a, 1957b, 1957c, 1958a, 1958b, 1958c, 1958d) and Comrey and Marggraff (1958). Among other things, these studies have indicated that the standard MMPI empirical scales are far from homogeneous in item content and that the dimensionality of the MMPI item pool might be greater than that suggested by factorial studies of the individual scales (Lingoes, 1960). It is important to note, however, that the substantive dimensions which emerged in these studies are dimensions which are defined in relation to the original empirical scales. These clusterings are based on only a portion of the total MMPI and represent subclusters of content "filtered through" the strategy of contrasted groups employed in the construction of the original scales. Such clusterings, no doubt, contain meaningful dimensions of item content, but, in addition, they contain variance peculiar to all dimensions along which the originally contrasted normal and psychiatric groups differed (Wiggins, 1962).

Attempts at more efficient measurement through factorially derived scales have also been conducted within the limited context of the original empirical scales. Welsh (1956) cluster analyzed nonoverlapping clinical scales to obtain markers for his item analytic procedures which yielded the well known *A* and *R* scales. Similarly, Eichman (1961, 1962) used the results of a factor analysis of clinical scales to derive his factor scales. Although working on the item level, Mees (1959) employed only 119 items selected from the standard clinical scales in developing his item factor scales. The fact that only subsets of items defined by the clinical scales were employed in these factorial studies probably does not seriously detract from their goal of more efficient measurement. Factor scales for the MMPI can be developed from almost any subsample of items (Wiggins & Lovell, 1965) and possibly from just a few direct statements (Peterson, 1965). Unfortunately, the factorial homogeneity of MMPI items has made the test particularly vulnerable to interpretations of stylistic (Edwards & Diers, 1962; Jackson & Messick, 1961; Messick & Jackson, 1961) and method (Wiggins & Lovell, 1965) components being involved or contaminated with substantive components. The extent of this contamination cannot be fully assessed until a serious effort has been made to illuminate the substantive dimensions of the total item pool rather than simply that portion of it which is most responsive to the strategy of contrasted groups.

ORIGINAL CONTENT CATEGORIES

In selecting items for possible inclusion in the final version of the MMPI, the "universe of content" (Loevinger, 1957) was deemed to be "behaviors of significance to the psychiatrist" (Hathaway & McKinley, 1940). With this in mind,

> ...the items were supplied from several psychiatric examination direction forms, from various textbooks of psychiatry, from certain of the directions for case taking in medicine and neurology, and from the earlier published scales of personal and social attitudes [Hathaway & McKinley, 1940, p. 249].

The names of the 26 categories suggested by the test authors as descriptive of item clusters in the MMPI pool are given in Table 1. To these 26 labels have been added phrases that are descriptive of the item content within each category.

Internal Consistency

As a first step in the investigation of the contribution of the original content categories to test variance, each category was considered to be a "scale" composed of *n* items which could be combined to yield a single total score for any individual. As a preliminary scoring method, each item was keyed in the direction of "deviance" as determined by the *infrequent* item option chosen in the Minnesota normal population (Hathaway & McKinley, 1951, pp. 26–29).[2] It should be emphasized that this scoring procedure is not entirely consistent with the empirically determined keying direction of the MMPI clinical scales, since several of the clinical scales contain items which are keyed in the popular direction. More important, such a scoring procedure in no way insures optimal scale homogeneity since both ends of an attitudinal continuum may be deviant with respect to population norms. In the case of sexual attitudes, for example, items admitting *both* antisexual attitudes and sexual acting out are deviant and hence both keyed in the same direction although such keying is intuitively inconsistent. However, in the absence of detailed information concerning such things as interitem correlations, the preliminary scoring method was considered the one most compatible with the original purpose of the item pool.

The internal consistency of content categories thus formed was assessed from the full-scale MMPI protocols of 500 Stanford University students in introductory psychology. Total scores on odd and even items within each of the 26 content categories were obtained separately for 250 men and 250 women students. Correlations between odd and even item totals, corrected by the Spearman-Brown formula for double test length, are given in Table 2.

The internal consistencies of the original content categories can be seen to vary from near-zero coefficients to coefficients in the .80s. Directly comparable internal consistency appraisals of the standard MMPI clinical scales have not been reported for college students taking the group form (Dahlstrom & Welsh, 1960, p. 474). However, the most comparable data available (Gilliland & Colgin, 1951) suggest that the majority of content categories have internal consistency coefficients equal to or greater than those reported for the standard MMPI clinical scales. As indicated in Table 2, the internal consistencies of many of the content categories are hampered by containing small numbers of items. Hathaway and McKinley (1940) were not explicit about the extent to which the number of items in a given category can be taken as representative of the relative significance of the category to a psychiatrist. Whether arising from implicit, explicit, or fortuitous circumstances, there are definite psychometric restrictions on the extent to which 5-item content categories may contribute to total test variance as contrasted with 55- or 72-item categories. When all content category internal consistency coefficients are corrected to the common base of the largest category (72 items), there is little to discourage an investigator from developing an expanded pool of items for any of the categories simply because some of them happen to be underrepresented in the MMPI. The obtained reliabilities are even more impressive in light of the previously

---

[2] It is important to note that such a keying procedure is correlated with "absolute" rather than "relative" deviance in Berg's (1961) usage. Although the infrequent item option in the Minnesota normals is often assumed to be an "abnormal" response (i.e., typical of psychiatric populations) such an assumption is more likely wrong than right (Ullmann & Wiggins, 1962).

TABLE 1

ORIGINAL CONTENT CATEGORIES OF THE MMPI

Affect, depressive (32 items): Sadness, despair, pessimism, futility; loneliness; guilt and expectation of punishment; worrying and brooding; sensitivity; anxiety; psychomotor retardation.

Social Attitudes (72 items): Introverted, seclusive; withdrawn; shy; nonoutgoing; non-fun-loving; overly sensitive; irritable; feels misunderstood; lacking in self-confidence; social rigidity; uncommunicative; lacking in social aggressiveness; critical and resentful of others.

Morale (33 items): Lacks self-confidence; low self-esteem; works and lives under tension; difficulties in concentrating, planning, making decisions, completing tasks; expects failure and resents success of others; suggestible and immature; feels misunderstood and unappreciated; sensitive and pessimistic.

Political Attitudes—law and order (46 items): Sees world as jungle, identification with criminal code, distrust of motives of others, discipline problem in school, delinquent childhood, thrill seeking, resentment and distrust of authority, competitive and vindictive, independence from norms, lack of concern for family members' misbehaviors, opinionated.

Obessessive and Compulsive states (15 items): Obsessions, compulsions, rumination, destructive impulses, covert defiance, overt compliance.

General Neurologic (19 items): Headaches, nausea, seizures, lability, poor judgment, distractibility, poor memory.

Vasomotor, Trophic, Speech, Secretory (10 items): Hot and cold sensations, sweating, blushing, dry mouth, poor reading comprehension.

Delusions, Hallucinations, Illusions, Ideas of Reference (31 items): Delusions of persecution and grandeur, ideas of influence, suspiciousness, hallucinations, bizarre experiences, malevolent forces in environment.

Phobias (29 items): Admission of general fearfulness and worry; specific irrational fears of animals, states of nature, disease, heights, crowds, etc.

Family and Marital (26 items): Lack of affection for parents; domination by parents; lack of parental support; desire to leave home; poverty; strife within family; disapproval, resentment, ambivalence, and annoyance at family members; disappointment in love; never been in love.

Lie Items (15 items): Naive and improbable claims to virtue with venial sins such as procrastination, vanity, gossip, citizenship, mild anger, bad thoughts, competitiveness, etc.

Masculinity-Femininity (55 items): Feminine interest pattern in literature, hobbies and childhood games; preference for feminine as opposed to masculine vocations; confused sexual identity; admission of weakness, fears, worries and distress.

General Health (9 items): Poor health, worry about health, high strung, weight fluctuation, easily tired.

Motility and Coordination (6 items): Muscular paralysis, contraction, tremor, weakness, and uncoordination.

Gastrointestinal System (11 items): Excessive and poor appetites, stomach trouble, constipation and diarrhea, lump in throat.

Affect, manic (24 items): Excitement, euphoria, high energy; restless and impulsive; irritability, quick temper, destructive impulses; optimism; flight of ideas; unpredictable; short memory; wide and short term interests; unusual hearing.

Occupational (18 items): Rigid work habits, distractability, sensitivity to opinions and criticisms of others, obstinance, indecisiveness, timidity, lack of self-confidence and concern about work, resentment of boss.

Cardiorespiratory System (5 items): Chronic cough, asthma or hay fever, chest pains, vomiting or coughing blood, pounding heart and shortness of breath.

Habits (19 items): Sleep disturbance, sensitivity to dreams, absence of dreams, excessive drinking and use of alcohol, abstinence from alcohol, giving in to bad habits.

Cranial Nerves (11 items): Disturbances in vision, speech, audition, olfaction and swallowing; facial paralysis.

Sensibility (5 items): Hypersensitivity to pain, touch; numbness; tingling skin sensations.

Educational (12 items): Dislike of reading—both fiction and nonfiction, likes funny papers and articles on crime, slow learner and disliked school.

Religious Attitudes (19 items): Fundamentalist beliefs, rejection of fundamentalist beliefs, unusual religious experiences, religiosity, magical beliefs, lack of praying and church attendance.

Sadistic, Masochistic Trends (7 items): Enjoys hurting and being hurt by loved ones, cruelty to animals, enjoys frightening people, fetishism.

Sexual Attitudes (16 items): Anxiety over sex, sexual preoccupation, sexual perversion, suppressive attitudes toward sex, permissive attitudes toward sex, disgust and embarrassment about sex.

Genitourinary System (5 items): Disturbance in urination, skin rash, something wrong with sex organs.

TABLE 2
CORRECTED ODD-EVEN RELIABILITY COEFFICIENTS FOR 26 ORIGINAL CONTENT CATEGORIES OF THE MMPI

Category	n	r_{xx} ($N = 250$ men)	r_{xx} ($N = 250$ women)
Affect, depressive	32	.865	.851
Social Attitudes	72	.850	.775
Morale	33	.802	.738
Political Attitudes	46	.727	.632
Obsessive-Compulsive	15	.668	.601
General Neurologic	19	.644	.693
Vasomotor	10	.632	.628
Delusions	31	.624	.470
Phobias	29	.622	.728
Family and Marital	26	.581	.471
Lie Items	15	.550	.652
Masculinity	55	.547	.505
General Health	9	.476	.307
Motility and Coordination	6	.475	.553
Gastrointestinal	11	.470	.353
Affect, manic	24	.454	.510
Occupational	18	.436	.396
Cardiorespiratory	5	.422	.477
Habits	19	.418	.588
Cranial Nerves	11	.417	.538
Sensibility	5	.338	.397
Educational	12	.333	.261
Religious Attitudes	19	.258	.184
Sadistic-Masochistic	7	.244	.302
Sexual Attitudes	16	.216	.249
Genitourinary	5	.169	.176
Total	550		

mentioned fact that the scoring procedures employed did not insure scale homogeneity.

It is of interest to note that the three content categories which have the highest internal consistencies for both men and women are Affect-Depressive, Social Attitudes, and Morale. As previously reported (Wiggins & Vollmar, 1959), these three categories account for some 70% of the item content of Welsh's *A* scale, the empirically derived marker of the potent first factor of the MMPI (Welsh, 1956). A decade ago, such an observation would have lent encouragement to a substantive interpretation of the first factor of the MMPI. It is now generally recognized that an unfortunate confounding of item characteristics and content mitigates against any such straightforward interpretation (Block, 1965; Dicken, Van Pelt & Bock, 1965; Wiggins, 1962; Wiggins & Goldberg, 1965).

Factorial Structure

Despite the fact that some of the original content categories are not reliably represented and that the present scoring method is less than optimal, it is of considerable interest to inquire into the number and kinds of substantive dimensions represented in the total MMPI item pool. Such an analysis would be the first to be based on a mutually exclusive and exhaustive classification of MMPI items.

Accordingly, product-moment intercorrelations among the 26 total scale scores were computed separately in the college samples of 250 men and 250 women. The matrices of content category intercorrelations were factored by the method of principal components (Harman, 1960). Latent roots which exceeded unity were retained and rotated analytically to a varimax criterion (Kaiser, 1958).

The method of analysis employed yielded seven factors[3] for men and six for women which accounted respectively for 60.9% and 55.1% of the total variance. The rotated factor matrices for men and women are presented in Tables 3 and 4. Factor loadings less than .33 have been omitted, and the matrices have been arranged in such a way as to facilitate comparison. Factor interpretation will be further facilitated by consulting the content category descriptions provided in Table 1.

Factor I appears to be the familiar general maladjustment dimension of the MMPI clinical scales. The content categories which load this factor most heavily reflect subjectively experienced distress on the part of the respondent. Low self-esteem, depressed mood, and feelings of inadequacy are coupled with social uneasiness and introversion. Anxiety is experienced directly with its usual physiological manifestations. This factor is loaded mod-

[3] Since communalities were arbitrarily set equal to unity, the resultant components are not to be considered "factors" in Thurstone's sense of this word, although they will be referred to as such.

TABLE 3
ROTATED FACTOR MATRIX OF 26 ORIGINAL CONTENT CATEGORIES
(N = 250 men)

	I	II	III$_a$	III$_b$	IV	V	VI	h^2
Morale	.74							.81
Vasomotor	.74							.66
Social Attitudes	.70			−.34				.68
Affect, depressive	.61	−.41		−.33				.82
Occupational	.45	−.53						.54
Phobias	.41		.39	−.39				.59
Affect, manic	.40	−.54						.58
Obsessive	.40	−.52						.58
Masculinity	.33	−.35				.44	−.54	.62
Political Attitude		−.63						.49
Sadistic		−.61						.53
Lie		.54						.61
Delusions		−.40	.39					.59
Sensibility		−.40	.53					.59
Cranial			.73					.61
Gastrointestinal			.55					.50
Genitourinary			.52		−.50			.56
General Neurologic			.50			.33		.63
Cardiorespiratory			.49	−.44				.51
Habits			.35			.45		.48
Motility				−.79				.69
General Health				−.68				.56
Sexual Attitudes					−.73			.64
Family and Marital					−.67			.62
Religious Attitudes						.74		.57
Educational							.85	.80
Variance (percent)	20.3	18.9	16.8	11.8	9.6	8.2	14.4	

erately by items reflecting irrational fears, restless irritability, and obsessional thinking. In men, this general maladjustment is also reflected in poor work habits and feminine interests. In women, maladjustment includes an unsatisfactory family background and a greater emphasis on poor physical health and undesirable habits.

Factor II is loaded by an intriguing combination of contents which have heretofore been observed to covary only in highly specialized instruments. The political attitudes category reflects authority conflict and authoritarian attitudes toward law and order. A sado-masochistic orientation is combined with obsessive-compulsive symptoms and overt compliance. Naïve and improbable claims to virtue may further suggest rigidity. Together these categories provide an almost classic description of the authoritarian personality syndrome which has been described in a variety of other contexts (Adorno et al., 1950; Loevinger, 1962; Rokeach, 1960; Stern, Stein, & Bloom, 1956). Although this factor is most clearly defined for women, the additional categories which load it for men are compatible with the areas of maladjustment often associated with authoritarianism. In men, the mood disturbances, poor work habits, sensitivities, delusional thinking, and feminine interests may reflect a more deep-seated personality disturbance than is the case with authoritarianism in women.

Two dimensions of physical symptoms in men (Factors IIIa and IIIb) appear to be combined in a single dimension of physical complaint for women (Factor III). Factor IIIa in men is loaded by a variety of complaints presumably representative of disturbance in cranial nerve, gastrointestinal, sensibility, genitourinary, neurologic,

TABLE 4
ROTATED FACTOR MATRIX OF 26 ORIGINAL CONTENT CATEGORIES
(N = 250 women)

	I	II	III	IV	V	VI	h^2
Morale	.75						.74
Vasomotor	.44						.52
Social Attitudes	.80						.70
Affect, depressive	.78			.34			.81
Occupational				.55			.52
Phobias	.41			.52			.56
Affect, manic	.41		−.41				.53
Obsessive	.47	.37	−.33				.59
Masculinity			−.33	.42		−.44	.57
Political Attitudes		.51	−.33		−.34		.58
Sadistic		.80					.66
Lie		−.35		−.56			.46
Delusions			−.57		−.34		.60
Sensibility			−.65				.52
Cranial			−.64				.43
Gastrointestinal			−.44				.44
Genitourinary			−.33			.44	.38
General Neurologic	.47		−.63				.63
Cardiorespiratory			−.56				.43
Habits	.37		−.38				.37
Motility			−.76				.71
General Health	.51		−.35				.50
Sexual Attitudes				.69			.50
Family and Marital	.54						.45
Religious Attitudes					−.84		.73
Educational						.63	.42
Variance (percent)	26.9	11.4	26.6	16.6	9.5	9.0	

and cardiorespiratory systems. Factor IIIb appears to center around general health concern, symptoms of fatigue and cardiorespiratory complaints. Both of these factors are loaded slightly by categories of psychological symptoms as well. In women, fatigue and general health concern are combined with the forementioned systemic symptoms into one general factor of somatic complaint. With the exception of the manic category reflecting fast tempo and irritability, the psychological symptom categories appear quite secondary in their contribution to this factor.

Factor IV is highly and uniquely loaded by the category of sexual attitudes. Categories associated with deviant sexual attitudes vary remarkably for men and women. In men such attitudes are associated with family conflict and specific genitourinary complaints. In women, this factor is negatively loaded by improbable claims to virtue and positively loaded by feminine interests. Psychological symptoms in women take the form of irrational fears, poor work habits, and feelings of depression and guilt.

Factor V is strongly and uniquely loaded by deviant religious attitudes. In men, sleep disturbance, drinking habits, and feminine interests tend to have loadings on this factor, and to a lesser extent, some somatic complaints. In women, authoritarian attitudes and delusional thinking have very slight loadings on the deviant religious attitudes factor.

Factor VI is defined by deviant educational attitudes. In both men and women, antieducational attitudes are associated with masculine interests. In women, this factor is also loaded by the category of genitourinary complaints.

In summary, a principal-component analysis of the 26 mutually exclusive and

exhaustive content categories of the MMPI yielded six interpretable factors in both men and women. The first three of these factors appear to represent general syndromes of complaint while the last three appear to center around more specific substantive categories. The factors of general maladjustment and somatic complaint are familiar ones which might be anticipated both on the basis of clinical scale development and of the overrepresentation of such categories in the MMPI item pool. The factor of authoritarianism appears to represent a theoretically meaningful combination of substantive categories which has, until now, been obscured by the strategy employed in the development of the clinical scales. Deviant attitudes towards sex, religion, and education have likewise not been previously stressed as important substantive components of the MMPI item pool.

Revision of Original Content Categories

Given the encouraging internal consistencies and factorial structure of the original content categories, it seemed fruitful to attempt a more substantively consistent grouping of items within categories as a basis for subsequent development of actual content *scales*. Although many strategies of scale construction were possible at this point, the one chosen placed primary emphasis on the "rational" or substantive considerations involved in the classification of item content. Since this strategy is so antithetical to the traditional approach to MMPI scale construction, a brief justification seems required.

For better or (more likely) for worse, the MMPI represents a *fixed* item pool. Examination of the interrelationships among many characteristics of this item pool led Wiggins and Goldberg (1965) to conclude:

> Over- and under-representation of certain classes of desirability, endorsement, ambiguity, and grammatical characteristics tends to make the item pool unnecessarily homogeneous and may, in part, contribute to rather severe restrictions in criterion group discriminations. The fortuitous confounding of such item characteristics with substantive dimensions (Block, 1965; Wiggins, 1962) has created interpretative problems (Edwards, 1957; Jackson & Messick, 1961) which may never be satisfactorily resolved within any fixed item pool [p. 394–395].

Although these authors stress the importance of basic research in item development, such research will not be of immediate value to the practical consumer of the MMPI. The present attempt to develop substantive scales for the MMPI was not initiated with the hope of overcoming the built-in shortcomings of the item pool. However, it was predicated on the assumption that the interaction of item characteristics and stylistic tendencies with substantive dimensions might be better understood than the interaction of such sources of variance with the complex and poorly understood dimensions yielded by the strategy of contrasted groups (e.g., "hysteria").

The method whereby the original MMPI content categories were revised involved the collapsing of several categories into single categories, reassignment of items from one category to another, elimination of original categories, creation of new ones and rekeying of item options within categories. Procedures were, with one minor exception, completely intuitive, and no claim is made for their replicability.[4]

The major item regroupings involved physical symptoms, interests, and items reflecting manifest hostility. Items from General Health, Cardiorespiratory, Gastrointestinal, and Genitourinary were combined in the single revised category of "Poor Health." Items from General Neurologic, Cranial Nerves, Motility and Coordination, and Sensibility were combined into a single category of "Organic Symptoms." Items reflecting hostility from the Sadistic-Masochistic category formed the nucleus of a new category of "Manifest Hostility" to which 21 items from 7 other original categories were added. A small

[4] The assistance of Victor R. Lovell in performing these item regroupings is gratefully acknowledged.

TABLE 5

CORRECTED ODD-EVEN INTERNAL CONSISTENCY COEFFICIENTS OF REVISED CONTENT CATEGORIES IN TWO COLLEGE POPULATIONS

Revised category	*n*	Stanford men (N = 250)	Stanford women (N = 250)	Oregon men and women (N = 203)
Religious	15	.87	.86	.81
Social	56	.84	.83	.80
Depression	33	.83	.82	.78
Morale	40	.84	.79	.74
Authority	43	.77	.71	.80
Phobias	27	.72	.80	.75
Hostility	27	.75	.73	.75
Organic	36	.71	.79	.72
Psychoticism	48	.75	.72	.71
Family	27	.74	.67	.73
Hypomania	25	.72	.71	.66
Health	28	.76	.70	.52
Feminine	56	.55	.60	.82
Sleep	15	.56	.58	.52
Obsessive	27	.52	.55	.50
Addiction	6	.67	.53	.42
Lie	15	.55	.65	.55
Vasomotor	10	.62	.60	.58
Sexual	16	.34	.51	.53

group of items from the Habits category were considered separately as an "Addiction" category.

The Occupational Attitudes category was judged too heterogeneous, and items from this category were regrouped under "Obsessive-Compulsive," "Poor Morale," and four other revised categories. Original categories that were retained were purified around a central theme and items eliminated or borrowed from other categories in light of this theme. The category of Habits, for example, was redefined as "Sleeping Habits" which eliminated seven of the original items and added three from other categories.

The categories of Educational Attitudes and Masculinity-Femininity were placed in a common pool and from this pool preliminary attempts were made to differentiate feminine interest patterns from tendencies toward sexual inversion. When this differentiation was judged unsuccessful, a general category of "Feminine Interests" was developed which proved to be ambiguous with respect to keying direction. In the absence of a clear-cut rationale, the empirical norms of Drake (1953) were used as a basis for item keying. Items in the "Feminine Interest" category which significantly differentiated men and women in Drake's sample were retained and keyed in the female direction.

Internal Consistency of Revised Categories

A more substantively consistent arrangement of items into content categories should be reflected in increased internal consistencies in the revised set. Total scores on odd and even items were computed for each of the 18 revised categories in samples of 250 men and 250 women students from Stanford University. Since these samples had, in part, inspired the reclassification, odd and even totals were also computed in a mixed group of 203 men and women introductory psychology students from the University of Oregon.[5] Table 5 presents Spearman-Brown corrected internal consistency coefficients for the Stanford and Oregon samples. Since new content categories were created and old ones considerably altered in the revision of the content categories, the success of the revision procedures cannot, in all cases, be directly assessed by comparison of each category with its revised counterpart. A slight decline in internal consistency occurred in depression, obsessive-compulsive and vasomotor categories. This is more than offset by the increases in internal consistency which occurred in 14 categories which can be compared with their original counterparts. The most dramatic increase occurred in the category of religious attitudes where *deletion* of four items and rekeying of those remaining resulted in internal consistency increases from the low 20s to the high 80s. Regrouping sadistic-masochistic items into the more general category of "Manifest Hostility" resulted in increases from the low .30s to the middle .70s. Other increases may be noted by comparing Table 5 with Table 2.

[5] These data for 95 men and 108 women were made available by Lewis R. Goldberg.

CONSTRUCTION OF FINAL CONTENT SCALES

On the basis of the data presented in Table 5, it was decided that there were 15 substantive dimensions in the MMPI pool which possessed promising internal consistencies and sufficient numbers of items to warrant further exploration. These 15 dimensions appear as the first 15 categories in Table 5. The categories of Addiction, Lie, Vasomotor, and Sexual were dropped from further consideration at this point. The categories of Feminine Interests, Sleeping Habits, and Obsessive-Compulsive were carried along on a very tentative basis.

The Stanford sample was randomly divided into two groups of 300 and 200 subjects, with an equal number of men and women within each group. The group of 300 subjects served as an item analysis group for scale purification and the group of 200 subjects was used for an independent assessment of the homogeniety of scales formed by item analysis.

Point biserial correlations were computed between the 566 items of the MMPI[6] and each of the 15 total scale scores of the revised content categories. An item was retained in a given content scale if: (*a*) its point biserial correlation with the total scale of the category of which it was a member exceeded .30 and (*b*) if its correlation with the total scale of the category of which it was a member exceeded its correlation with all 14 remaining revised content category scores.

Table 6 shows the number of items eliminated by each of the two criteria of item analysis. Among the Social Maladjustment items, for example, 26 items were eliminated because their correlation with the Social total scale score was less than .30. Three additional items were eliminated because their item-total correlations, although greater than .30, were equaled or exceeded by item-total correlations with one or more of the 14 additional content categories.

TABLE 6

NUMBER OF ITEMS ELIMINATED BY ITEM ANALYSIS OF REVISED CONTENT CATEGORIES

Scale	Original	$r < .30$	Nonindependent	Final n
Religious	15	3	0	12
Social	56	26	3	27
Depression	33	12	2	20[a]
Morale	40	15	3	23[a]
Authority	43	21	2	20
Phobias	27	7	1	19
Hostility	27	7	1	20[a]
Organic	36	13	0	23
Psychoticism	48	31	4	13
Family	27	11	0	16
Hypomanic	25	5	0	20
Health	28	15	0	13
Feminine	56	26	0	30
Sleep	15	3	1	11
Obsessive	27	14	3	10

[a] Includes one additional item from another content category.

It can be seen from Table 6 that item selection was made primarily on the basis of internal consistency. Only 20 items were eliminated on the basis of their being correlated with categories other than their own. Note, however, that the criterion of scale independence employed was quite minimal. Three items were judged to have been initially misclassified after examination of their correlations with other categories. Thus, one "Obsessive" item was transferred to the Depression category, one "Social" item to the Morale category, and one "Psychoticism" item to the Hostility category.

The 15 revised content categories and the 15 content scales formed by item analysis were then scored in the group of 200 subjects originally set aside for this purpose. As a more general measure of scale homogeneity, Cronbach's coefficient alpha (Cronbach, 1951) was computed for the 15 categories and 15 scales. Content scales were judged improved by item analysis if their alpha coefficient increased despite the elimination of substantial proportions of items.

Table 7 presents alpha coefficients for

[6] Sixteen items are repeated in the group form of the MMPI. In all analyses reported here, only the first appearance of a repeated item was considered.

TABLE 7

COEFFICIENT ALPHA INTERNAL CONSISTENCY ESTIMATES FOR REVISED CATEGORIES AND ITEM-ANALYZED CONTENT SCALES
(N = 200 men and women)

Category	Revised	n	Final	n	Revised versus final
Religious	81	(15)	83	(12)	98
Social	83	(56)	86	(27)	95
Depression	84	(33)	82	(20)	96
Morale	81	(40)	84	(23)	93
Authority	77	(43)	78	(20)	92
Phobias	70	(27)	67	(19)	96
Hostility	72	(27)	69	(20)	97
Organic	76	(36)	70	(23)	96
Psychoticism	76	(48)	61	(13)	85
Family	72	(27)	72	(16)	94
Hypomanic	69	(25)	67	(20)	97
Health	69	(28)	59	(13)	90
Feminine	77	(56)	84	(30)	96
Sleep	56	(15)	56	(11)	97
Obsessive	56	(27)	57	(10)	82

the revised categories and the scales formed by item analysis. The contaminated correlation between the two sets of measures is presented in the final column. The Religion, Social, Morale, Authority, Family, and Feminine Interests scales were judged to be improved by item analysis. Scale purification was extreme in several instances and resulted in improved alphas despite elimination of almost half the items in the scale.

Increased homogeneity was not achieved by item analysis for Depression, Phobias, Hostility, Organic, Psychoticism, Hypomania, or Health. Subsequent attempts to improve these scales by less stringent item analytic criteria were not successful. It was decided, therefore, to retain these scales in their revised form. The Sleeping Habits and Obsessive scales were abandoned at this point on the grounds of unpromising homogeneity.

The foregoing procedures resulted in the adoption of 13 mutually exclusive scales which were considered to be internally consistent, moderately independent, and representative of the major substantive clusters of the MMPI. All of these scales were based on rational regroupings of the original content categories proposed by Hathaway and McKinley. Six of these scales were further refined by item-analytic procedures. This final set of 13 scales will be referred to as the MMPI content scales.[7] The content of the items in the scales is described in Table 8.

Internal Consistency of Content Scales in Normal Populations

Since virtually all of the preliminary investigation and development of the MMPI content scales was based on a single college population, it was necessary to gather additional data from other populations to assess the psychometric characteristics of the final scales. Accordingly, complete MMPI protocols were obtained from the samples listed in Table 9. A group of Air Force enlisted men served as a noncollege normal population while the remaining samples were college students of both sexes from several geographical regions.[8]

The internal consistency of the MMPI content scales was assessed by computing alpha coefficients in samples not involved in scale derivation. These data are presented in Table 10. Reliability coefficients from the college samples are, with one notable exception, generally in accord with expectations gained from the derivation samples. The exception is Feminine Interests, which, although among the most internally consistent scales in the derivation sample, is the least reliable scale in other college and Air Force samples. More in line with expectations are the generally high internal consistencies of Social Maladjustment, Religious Fundamentalism, Depression, and Poor Morale in the college groups. As before, Hypomania and Poor Health are among the lowest in internal consistency, but the obtained alpha coefficients are quite respectable in comparison with the majority of MMPI scales in use today.

With the exception of Feminine Interests, the alpha coefficients obtained in

[7] Item lists are provided in Appendix A.

[8] The author is grateful to John D. Hundleby and to Leonard G. Rorer for making available the Air Force and Minnesota college data, respectively.

TABLE 8

DESCRIPTION OF MMPI CONTENT SCALES

SOC *Social Maladjustment:* High SOC is socially bashful, shy, embarrassed, reticent, self-conscious and extremely reserved. Low SOC is gregarious, confident, assertive, and relates quickly and easily to others. He is fun loving, the life of a party, a joiner who experiences no difficulty in speaking before a group. This scale would correspond roughly with the popular concept of "introversion-extraversion."

DEP *Depression:* High DEP experiences guilt, regret, worry, unhappiness and a feeling that life has lost its zest. He experiences difficulty in concentrating and has little motivation to pursue things. His self-esteem is low, and he is anxious and apprehensive about the future. He is sensitive to slight, feels misunderstood, and is convinced that he is unworthy and deserves punishment. In short he is classically depressed.

FEM *Feminine Interests:* High FEM admits to liking feminine games, hobbies, and vocations. He denies liking masculine games, hobbies, and vocations. Here there is almost complete contamination of content and form which has been noted in other contexts by several writers. Individuals may score high on this scale by presenting themselves as *liking* many things since this item stem is present in almost all items. They may also score high by endorsing interests, which, although possibly feminine, are also *socially desirable* such as an interest in poetry, dramatics, news of the theatre, and artistic pursuits. This has been noted in the case of Wiggins' *Sd* scale. Finally, of course, individuals with a geniune preference for activities which are conceived by our culture as "feminine" will achieve high scores on this scale.

MOR *Poor Morale:* High MOR is lacking in self-confidence, feels that he has failed in life and is given to despair and a tendency to give up hope. He is extremely sensitive to the feelings and reactions of others and feels misunderstood by them while at the same time being concerned about offending them. He feels useless and is socially suggestible. There is a substantive overlap here between the Depression and Social Maladjustment scales and the Poor Morale scale. The Social Maladjustment scale seems to emphasize a lack of social ascendance and poise, the Depression scale feelings of guilt and apprehension, while the present scale seems to emphasize a lack of self-confidence and hypersensitivity to the opinions of others.

REL *Religious Fundamentalism:* High scorers on this scale see themselves as religious, church-going people who accept as true a number of fundamentalist religious convictions. They also tend to view their faith as the true one.

AUT *Authority Conflict:* High AUT sees life as a jungle and is convinced that others are unscrupulous, dishonest, hypocritical, and motivated only by personal profit. He distrusts others, has little respect for experts, is competitive and believes that everyone should get away with whatever they can.

PSY *Psychoticism:* High PSY admits to a number of classic psychotic symptoms of a primarily paranoid nature. He admits to hallucinations, strange experiences, loss of control, and classic paranoid delusions of grandeur and persecution. He admits to feelings of unreality, daydreaming, and a sense that things are wrong, while feeling misunderstood by others.

ORG *Organic Symptoms:* High ORG admits to symptoms which are often indicative of organic involvement. These include headaches, nausea, dizziness, loss of motility and coordination, loss of consciousness, poor concentration and memory, speaking and reading difficulty, muscular control, skin sensations, hearing and smell.

FAM *Family Problems:* High FAM feels that he had an unpleasant home life characterized by a lack of love in the family and parents who were unnecessarily critical, nervous, quarrelsome, and quick tempered. Although some items are ambiguous most are phrased with reference to the parental home rather than the individual's current home.

HOS *Manifest Hostility:* High HOS admits to sadistic impulses and a tendency to be cross, grouchy, competitive, argumentative, uncooperative, and retaliatory in his interpersonal relationships. He is often competitive and socially aggressive.

PHO *Phobias:* High PHO has admitted to a number of fears, many of them of the classically phobic variety such as heights, dark, closed spaces, etc.

HYP *Hypomania:* High HYP is characterized by feelings of excitement, well being, restlessness, and tension. He is enthusiastic, high strung, cheerful, full of energy, and apt to be hotheaded. He has broad interests, seeks change, and is apt to take on more than he can handle.

HEA *Poor Health:* High HEA is concerned about his health and has admitted to a variety of gastrointestinal complaints centering around an upset stomach and difficulty in elimination.

the Air Force sample are substantial, indicating a generality beyond college populations. Several differences in the relative internal consistencies of the content scales in an Air Force, as opposed to college, population may be noted. Whereas Psychoticism and Organic Symptoms are only moderately reliable in college groups,

TABLE 9
COMPOSITION OF NORMAL SAMPLE
(N = 1,368)

Group	Men	Women
Air Force enlisted men[a]	261	—
Stanford University	250	250
University of Minnesota	96	125
University of Oregon	95	108
University of Illinois	100	83
	802	566

[a] Chanute Air Force Base, Rantoul, Illinois.

they are among the most internally consistent scales in the Air Force sample. This may reflect, in part, the greater heterogeneity of the Air Force sample. It is also of interest to note that whereas Religious Fundamentalism is consistently among the most reliable scales for college groups, it is one of the least reliable scales in the Air Force sample.

TABLE 10
COEFFICIENT ALPHA INTERNAL CONSISTENCY ESTIMATES FOR MMPI CONTENT SCALES IN SEVEN NORMAL SAMPLES

	AF enlisted men (N = 261)	University of Minnesota men (N = 96)	University of Minnesota women (N = 125)	University of Oregon men (N = 95)	University of Oregon women (N = 108)	University of Illinois men (N = 100)	University of Illinois women (N = 83)
SOC	.829	.856	.835	.830	.862	.856	.843
DEP	.872	.860	.831	.821	.756	.842	.854
FEM	.585	.523	.505	.594	.566	.650	.542
MOR	.857	.866	.825	.804	.753	.867	.804
REL	.674	.892	.861	.842	.756	.817	.793
AUT	.681	.794	.772	.743	.669	.766	.698
PSY	.877	.794	.687	.738	.662	.763	.806
ORG	.863	.772	.645	.652	.695	.749	.731
FAM	.707	.712	.789	.712	.694	.806	.643
HOS	.764	.819	.794	.788	.651	.776	.765
PHO	.765	.663	.721	.568	.701	.705	.770
HYP	.671	.701	.715	.682	.632	.679	.667
HEA	.743	.557	.713	.555	.537	.673	.651

Group Differences in Content Scale Scores

Personality inventory scale scores which presumably reflect individual differences along dimensions of substantive interest should, at the very least, be expected to reflect such differences when diverse groups are compared. The standard MMPI clinical scales were *constructed* to reflect differences among certain groups, and, initially, substantive interpretations of clinical scales were rather narrowly restricted to such differences, whatever they may imply. By contrast, the MMPI content scales were designed to reflect reliable individual differences along interpretable substantive dimensions and group differences, where found, will serve to enhance rather than define the meaning of the content scale involved.

The cooperation of two quite different psychiatric installations was obtained in securing complete MMPI protocols of patients on whom a final psychiatric diagnosis had been made.[9] One installation was a large state mental hospital whose inmates represent a wide spectrum of psychopathology, the most frequent diagnosis being that of chronic schizophrenia. The second installation was an outpatient clinic attached to an Air Force base whose clientele consists primarily of neurotic, sociopathic, and personality disorders. At each installation, an attempt was made to obtain the majority of recent and complete MMPI protocols on patients whose files were sufficiently complete to allow de-

[9] The author is indebted to Paul Finkel, Clifford M. Broadway, and other staff of Kankakee State Hospital for their assistance in providing protocols and case folders.

TABLE 11
COMPOSITION OF PSYCHIATRIC SAMPLE
(N = 614)

APA code	Diagnosis	Inpatients[a]		Out patients[b]	
		Men	Women	Men	Women
000–199	Brain disorders	23	16	16	—
200–213	Affective psychoses	20	27	—	—
220–229	Schizophrenic psychoses	85	83	—	4
400–406	Psychoneurotic disorders	13	23	15	7
500–504	Personality pattern disturbance	15	5	17	2
510–513	Personality trait disturbance	17	8	36	6
520–524	Sociopathic personality disturbance	46	14	19	1
530–535	Special symptom reaction	—	—	6	—
540–546	Transient situational disturbance	—	—	8	3
	Other[c]	53	16	8	2
	Total	272	192	125	25

[a] Kankakee State Hospital, Kankakee, Illinois.
[b] Chanute Air Force Base Outpatient Clinic, Rantoul, Illinois.
[c] Rare category or indeterminant diagnosis.

termination of the final psychiatric diagnosis. On the basis of information contained in the case folder, each patient was classified in terms of the first three digits of the diagnostic code given in the American Psychiatric Association's (1952) *Diagnostic and statistical manual: Mental disorders.* In the inpatient sample, several preliminary diagnostic impressions were available in addition to the final, official hospital diagnosis made by the diagnostic staff. Where there was great discrepancy between preliminary and final diagnoses or where the final diagnosis was lacking in precision, the case was classified as "indeterminate." In the outpatient sample, only the final decision of the diagnostic staff was employed, and when this was imprecise, an "indeterminant" classification was assigned. The distribution of such classifications is given in Table 11 for both inpatient and outpatient samples. These distributions represent available records rather than any attempted sampling procedure. They are judged to be reasonably representative of the two kinds of installations involved.

Although the MMPI is given, more or less, routinely at both of these installations, its contribution to final psychiatric diagnosis is probably less than at other installations that routinely give the MMPI. It should be recognized, nevertheless, that an unknown degree of criterion contamination exists. However, in no instance were MMPI content scale scores available to the institution prior to final diagnosis.

From the samples listed in Tables 9 and 11, it was possible to form seven fairly large groups which differed markedly among themselves in such characteristics as age, sex, education, and psychiatric status. These groups were: (*a*) AFM (261 Air Force men); (*b*) IPM (272 inpatient men); (*c*) IPW (192 inpatient women); (*d*) OPM (125 outpatient men); (*e*) OPW (25 outpatient women); (*f*) CM (96 University of Minnesota college men); and (*g*) CW (125 University of Minnesota college women). The 13 MMPI content scales were scored in each of these seven groups. For each content scale, a simple analysis of variance was computed to test the null hypothesis that mean scale scores are the same for the populations from which the seven groups are derived. This hypothesis was rejected for all 13 scales at $p < .01$ by the F ratio with 6 and 1089 degrees of freedom. Differences between certain group means in content scales were further assessed by t tests for independent groups. Of 21 possible group comparisons, 13 were judged sensible, and a highly condensed summary of

TABLE 12
CONFIDENCE LEVELS FOR t TESTS OF GROUP MEAN DIFFERENCES IN CONTENT SCALE SCORES

	SOC	DEP	FEM	MOR	REL	AUT	PSY	ORG	FAM	HOS	PHO	HYP	HEA
AFM versus OPM	.05	.01	.01	.01	.001			.01					
OPM versus OPW	.05	.001	.001	.001		.05	.05	.001		.001	.001		.001
OPW versus CW	.001	.001		.001		.01	.001	.001	.001	.001	.001	.05	.001
CW versus IPW	.001	.001	.05	.001	.05	.001	.001	.001	.001	.001	.001		.001
OPM versus CM	.01	.001		.001		.001	.001	.001	.01		.001	.05	.001
IPW versus OPW	.01	.001		.001				.001		.05	.001		.001
OPM versus IPM	.05	.01	.001	.01	.05			.01		.05			
CM versus AFM	.05	.001		.001	.001	.001	.001	.001	.05	.001	.001	.001	.001
CM versus IPM		.001	.001	.001		.001	.001	.001	.01		.001	.001	.001
AFM versus IPM			.001		.01					.001		.05	
CM versus CW			.001	.01	.05	.001				.001	.001		
F ratio	393.14	23.18	255.76	20.43	8.20	33.21	17.95	34.71	436.46	14.31	32.64	2.94	25.27

these 13 group comparisons is presented in Table 12.

The last row of Table 12 contains the F ratios for each content scale based on all seven groups. The remaining rows contain the significance levels at which the hypothesis of no difference in underlying population means is rejected using t tests for independent means. In the first row, for example, means on the 13 content scales were compared for the Air Force men and the outpatient men groups. The hypothesis of no difference in group means was rejected at $p < .05$ for Social Maladjustment, at $p < .01$ for Depression, and at $p < .001$ for Religious Fundamentalism. Significant mean differences on the Authority Conflict scale were *not* found between Air Force men and outpatient men. This type of summary does not provide information on the extent of mean differences nor even their direction. Also, with this many comparisons it is to be expected that at least several will not be replicable. The point of the present analysis is not to attach significance to any single comparison but rather to provide an overview of the content scales which differ most from sample to sample and of the sample comparisons which yield the greatest differences.

The content scales which contributed most to differences among this particular sample of diverse groups were Poor Morale, Organic Symptoms, Phobias, and Depression. Lesser, but not insubstantial, differences occurred with Poor Health, Manifest Hostility, Feminine Interests, Authority Conflict, and Psychoticism. Content scales whose means did not differ greatly among the present samples are Social Maladjustment, Religious Fundamentalism, Family Problems, and Hypomania.

As might be expected, mean content scale scores differ most when college groups are compared with same-sex patient groups, both inpatient and outpatient. What might not be anticipated are the substantial differences between college and Air Force men. The differences between outpatient men and women probably reflect the un-

TABLE 13
CONFIDENCE LEVELS FOR t TESTS OF GROUP MEAN DIFFERENCES IN CLINICAL SCALE SCORES

	L	*F*	*K*	*Hs*	*D*	*Hy*	*Pd*	*Mf*	*Pa*	*Pt*	*Sc*	*Ma*	*Si*
AFM versus OPM				.001	.001	.001	.001			.001			.001
OPM versus OPW		.05	.01	.001	.001	.001		.001	.01	.001	.01		.001
OPW versus CW	.001	.001	.001	.001	.001	.001	.001	.05	.001	.001	.001	.01	.001
CW versus IPW	.001	.001	.001	.001	.001		.001	.001	.001	.001	.001	.001	.001
OPM versus CM	.001	.001	.001	.001	.001	.001	.001		.001	.001	.001	.01	.001
IPW versus OPW		.01	.01	.001	.001	.001	.01		.01	.001	.001		.001
OPM versus IPM				.001	.001	.001			.001	001	.01		.01
CM versus AFM	.001	.001	.001	.001	.001		.001	.01	.01	.001	.001	.001	.001
CM versus IPM	.001	.001	.05	.001	.001		.001	.05	.001	.01	.001		.001
AFM versus IPM		.05			.001		.001				.05	.001	
CM versus CW	.05	.01				.001	.001	.001					.001
F ratio	21.62	16.56	8.81	35.55	40.61	27.20	31.99	200.71	10.23	20.71	21.50	9.69	20.5

representative nature of a female sample at an Air Force installation. The Air Force "normal" sample differs very slightly from both inpatient and outpatient male samples. Differences between outpatient and inpatient men are also slight.

To provide a context of comparison for group differences obtained with the content scales, the same analysis was performed using the 13 standard MMPI clinical scales. A summary of this analysis is presented in Table 13. The number of significant differences obtained with the clinical scales is similar to that obtained with the content scales. Again, the greatest mean scale score differences occur when college groups are contrasted with patient groups. Relatively few differences are found between Air Force and patient samples or between inpatient and outpatient males. Although the number of significant differences is similar, the clinical scales in general tend to allow for rejection of the null hypothesis at a slightly higher significance level than is possible with the content scales. Considering that the clinical scales were specifically constructed for the purpose of discriminating normal from abnormal samples, the slight edge they possess over the content scales in the present analysis is not an impressive one. Whatever is represented in mean scale score comparisons across diverse groups is clearly present in the MMPI content scales as well.

Ordering of Group Means on Content Scales

In addition to assessing the overall contribution of content scale scores to group differences, it is of interest to examine the ordering of means for diverse groups within each of the separate content scales. Such a procedure is useful in suggesting underlying psychological continua which may be associated with scale scores (Gough, 1960). In the present instance this approach should not be considered validational as the content scales were not devised to serve specific group discriminative purposes. It is assumed that the dimensions underlying the content scales

are substantive and although primarily pathological in nature not necessarily equivalent to the dimensions which contribute to the fact of membership in a socially or psychiatrically defined group.

From the male samples described in Tables 9 and 11, 16 subgroups were selected to represent a broad range of socioeconomic, educational, and psychiatric variables. Content scale means and standard deviations for each subgroup are provided separately for the 13 content scales in Appendix B (Tables B1–B13). For each content scale, the groups have been ordered by mean scale score. In addition to providing suggestions of underlying dimensions, the data in Appendix B provide preliminary and admittedly inadequate normative information for male samples.

Even a cursory inspection of the tables in Appendix B reveals that the content scales, as a group, do not provide a measure of "pathology" that is consistent with the conventional psychiatric meaning of this term. Space limitations do not permit a detailed scale-by-scale analysis, but certain generalizations may be stated. When the rank order of each group within each scale is pooled across the 13 scales, four reasonably consistent groupings may be distinguished. In the first group are the brain disorders, the outpatient personality pattern and trait disturbances, and the Air Force normals—all of whom tend to be among the highest scorers on the content scales. Next in order are the inpatient and outpatient sociopathic disturbances, the affective psychoses, and the special symptom outpatient group. Following this group are the inpatient and outpatient neurotic disorders, the schizophrenic psychoses, and the inpatient personality trait disturbances. The lowest scoring group tends to consist of the inpatient personality trait disturbances, the outpatient brain disorders, college students, and outpatient transient situational disturbances. The foregoing generalizations tend to obscure large individual differences among content scales. The pattern of such differences can only be discerned by detailed examination of Tables B1 through B13.

Differential Diagnosis of Psychiatric Inpatients

The preceding consideration of scale mean distributions across a variety of samples was designed to explicate the meaning of MMPI scales formed from substantive rather than group discriminative considerations. Should such scales be applied to problems of psychiatric classification, it is hoped that the approach would be multivariate rather than single scale. Further, the problem typically facing the practicing diagnostician is not that of distinguishing disparate groups such as Air Force personnel and college students, but rather that of distinguishing putative subgroupings within a single population.

As an example of the use of MMPI content scales in a realistic diagnostic problem, multiple discriminant analytic procedures were applied to the classification of psychiatric inpatients. From Table 11 it can be seen that by combining the diagnostic categories of personality pattern and personality trait disturbance into the single category of personality disturbance (men = 32; women = 13) and by eliminating the rare and indeterminate categories (men = 53; women = 16), six major diagnostic groupings can be formed for men ($n = 219$) and women ($n = 176$), respectively. These groupings are: (*a*) brain disorders, (*b*) affective psychoses, (*c*) schizophrenic psychoses, (*d*) psychoneurotic disorders, (*e*) personality disorders, and (*f*) sociopathic disorders.

Using the 13 MMPI content scales as predictors, multiple discriminant analyses were performed separately on these six groups of men and six groups of women inpatients. The main purpose of this analysis was to test the generalized, multivariate null hypothesis that these six diagnostic groups have similar content scale scores. Should rejection of this hypothesis seem tenable, the contribution of the separate content scales to the main discriminant functions would shed light on their relative diagnostic importance. Evaluation of the replicability of the functions derived and

TABLE 14
DISCRIMINANT ANALYSIS OF SIX INPATIENT DIAGNOSTIC GROUPS BASED ON 13 MMPI CONTENT SCALES
(N = 219 men)

Scales	Scaled vectors				
	I	II	III	IV	V
REL	−.058	−.046	−.003	−.098	.082
SOC	−.058	.047	.239	−.032	−.105
DEP	.060	−.727	−.146	.114	.056
MOR	−.229	.300	.040	−.182	.508
AUT	−.402	.016	−.210	.196	−.080
PHO	−.054	.220	−.053	.291	−.004
HOS	.518	.054	−.002	−.342	−.213
ORG	−.187	−.293	−.087	−.040	−.354
PSY	−.134	.200	.573	.040	−.037
FAM	.117	.006	.016	.260	.091
HYP	−.091	.107	−.465	−.140	−.232
HEA	.145	.132	.091	−.206	.306
FEM	−.004	.037	.287	.007	−.067

Latent roots	Trace (percent)
$\lambda 1$ = .2089	40.62
$\lambda 2$ = .1522	29.60
$\lambda 3$ = .0894	17.38
$\lambda 4$ = .0525	10.20
$\lambda 5$ = .0113	2.19

Trace = .5142; Λ = .6192; $F^{65}_{954} = 1.56$, $p < .01$.

their efficiency in classifying other samples of psychiatric patients must await further data collection. The method of analysis and presentation of findings is that of Cooley and Lohnes (1962, pp. 116–133).

Tables 14 and 15 present a summary of findings from the discriminant analysis of six groups based on 13 content scales for the 219 men and 176 women, respectively. Where the number of groups is less than the number of predictors, the maximum number of discriminants is one less than the number of groups or, in the present instance, five. The five latent roots are presented along with their associated vectors (coefficients) which have been adjusted to permit comparison of their relative contribution to the discriminant function. The generalized null hypothesis was evaluated by Wilks' lambda which expresses the ratio of pooled within groups cross-product deviation scores to total sample cross-product deviation scores. In testing the significance of lambda, the F approximation of Rao was employed (Cooley & Lohnes, 1962, pp. 61–62). In the male sample, $F = 1.56$, which for 65 and 954 df is significant at $p < .01$. For the women $F = 1.53$, which for 65 and 750 df is also significant at $p < .01$. This makes tenable the rejection of the hypothesis of the equality of mean vectors for the six groups.

The coefficients associated with each of the five discriminant functions are of interest in assessing the relative contributions of the content scales to classification of an inpatient population. From a practical standpoint, it should be noted that three discriminant functions are probably sufficient for both men and women as they account for approximately 88% of the discriminating power of the scales in each sample. It should also be noted that the three discriminant functions in each analysis are sufficiently different in pattern for men and women as to discourage pooling data for sexes in a hospital population.

For men the largest contributors to group discrimination along the first discriminant function are Hostility and Authority Con-

TABLE 15
DISCRIMINANT ANALYSIS OF SIX INPATIENT DIAGNOSTIC GROUPS BASED ON 13 MMPI CONTENT SCALES
(N = 176 women)

Scales	Scaled vectors				
	I	II	III	IV	V
REL	.018	.082	.044	−.022	.042
SOC	.090	−.108	−.056	.097	.165
DEP	−.056	−.187	−.325	−.527	−.137
MOR	.377	.299	.086	.189	.211
AUT	.774	−.173	.381	−.090	−.129
PHO	−.159	.046	.010	−.126	.198
HOS	−.196	.021	−.304	.065	−.253
ORG	.257	.223	−.024	.186	−.316
PSY	−.406	−.055	.020	.120	.149
FAM	.104	−.230	−.104	.166	−.118
HYP	.062	−.005	.036	.008	.159
HEA	−.218	−.049	.167	.182	.224
FEM	.035	−.179	.015	.125	.082

Latent roots	Trace (percent)
$\lambda 1$ = .3758	55.54
$\lambda 2$ = .1382	20.43
$\lambda 3$ = .0834	12.33
$\lambda 4$ = .0488	7.21
$\lambda 5$ = .0304	4.49

Trace = .6765; Λ = .5455; F^{65}_{760} = 1.53, $p < .01$.

flict. Inspection of group means for these scales (Tables B10 and B6 of Appendix B) indicates that they are relatively effective in separating sociopathic and brain-disorder groups from schizophrenic, neurotic, and personality disturbances. Depression and Poor Morale contribute to group discrimination along the second discriminant. Mean Depression scale scores (Table B2) for brain disorder and neurotic groups are well above those for personality, affective and schizophrenic groups. Mean scores for Poor Morale (Table B4) reflect a separation between the brain-disorder group and the others just mentioned. Psychoticism and Hypomania are the largest contributors to the third discriminant function. Mean PSY scores (Table B7) suggest a clear psychotic-neurotic distinction with brain, sociopathic, schizophrenic, and affective groups in the former category and neurotic and personality groups in the latter. Mean HYP scores suggest a similar dichotomy with the interesting exception of schizophrenics being classified toward the "neurotic" pole of the implied continuum.

In the analysis based on women (Table 15) the largest contributor to discrimination along the first discriminant function is Authority Conflict. Inspection of group means for this scale indicates, as with the men, a separation of sociopathic ($\bar{X}$ = 10.71) and brain disorder (9.50) groups from schizophrenics (8.96) neurotics (8.52) and personality groups (8.75, 8.20). The affective psychosis group attained the lowest mean score on this scale (7.63). The Psychoticism scale is the second largest contributor to the first discriminant function. In women, high mean PSY scores are obtained for schizophrenics (13.78) and sociopaths (10.43), while lower mean scores characterize neurotics (9.57), brain disorder (8.44), affective (7.74), and personality groups (9.40, 7.13). Poor Morale contributes additionally to the first discriminant function and has the largest coefficient on the second. High MOR scores were obtained

for personality pattern (12.60), neurotic (12.16), and affective psychotic (10.89) groups. Lower scores were obtained by personality trait (9.88), sociopaths (9.07), and brain disorders (8.50).

Depression and Hostility (which were significant contributors to the first and second discriminants for men) contribute, in addition to Authority Conflict, to the third discriminant function for women. Mean DEP scale scores for women form a continuum on which personality pattern (12.80), neurotic (11.39), and schizophrenic (10.78) groups are high and affective (9.89) and brain disorder (6.69) groups are low. On the Hostility scale, schizophrenic (9.73) and personality groups (8.60, 8.37) are high, while brain disorder (7.25) and affective psychotic (7.22) groups are relatively low.

Any attempt to summarize the relative importance of the content scales in contributing to classification of psychiatric patients by multiple discriminant analysis must be restricted in generalization to the present sample. In addition to possible problems of sample specificity, the present analysis was restricted to six diagnostic groupings which, although not arbitrary, may not be the groupings desired in other hospital settings. Nevertheless it seems important to note that Authority Conflict, Poor Morale, Hostility, Psychoticism, and Depression were important contributors to group discrimination as were, to a lesser extent, Family Problems, Organicity, and Hypomania. Scales which contributed relatively little to the present analysis were Religious Fundamentalism, Social Maladjustment, Phobias, Poor Health, and Feminine Interests.

Discriminant Analysis Based on MMPI Clinical Scales

Multiple discriminant analyses were also performed using the 13 standard MMPI clinical scales as predictors of the six diagnostic groupings. Although not the primary concern of the present study, it was hoped that such analyses would provide a context of comparison for the analyses of content scales as well as insight into the utility of a multivariate approach to this most familiar diagnostic problem. Summaries of the results of these analyses for men and women are provided in Tables 16 and 17.

As before, the significance of Wilks' lambda was evaluated by the F approximation of Rao. In the male sample, $F = 1.62$ which for 65 and 954 df is significant at $p < .01$. For the women, $F = 1.49$ which for 65 and 750 df allows rejection of the null hypothesis at $p < .01$. As with the content scales, the hypothesis of no difference between mean vectors for the six groups can be confidently rejected. Again, from a practical standpoint the dimensionality of the predictor space might be reduced to three since the first three discriminant functions account for 86% and 88% of the discriminating power of the scales in male and female samples, respectively.

In the male sample, the predominant contribution of *Sc* and *Pt* to group discrimination can be seen to be operative in all but the fourth discriminant function. *Hy* and *K* contribute additionally to the first discriminant while *Si* and *Pd* are of importance to the second. The contribution of *F* to group discrimination is seen in the third and fourth discriminant. *Pa* is involved in the last three discriminants and *Hs* contributes to the fourth. Clinical scales *L*, *D*, *Mf*, and *Ma* contributed relatively little to the present analysis.

The first discriminant function in the female sample is even more clearly dominated by *Sc* and *Pt*; the latter scale in this instance being the more heavily weighted. *Hy* contributed additionally to the second discriminant and with *Pd* to the third as well. The scales which contribute most to the present discriminant analysis are clearly *Pt*, *Sc*, *Hy*, and *Pd*. Lesser contributions come from *F*, *Hs*, *D*, *Ma*, and *K*. Scales *L*, *Mf*, and *Si* are of only minor importance.

FACTORIAL STRUCTURE OF CONTENT SCALES

Unlike the standard MMPI clinical scales, the MMPI content scales do not share common items and were constructed in such a way as to maximize the homogeneity of each scale. Nevertheless, the criterion

TABLE 16
DISCRIMINANT ANALYSIS OF SIX INPATIENT DIAGNOSTIC GROUPS BASED ON 13 MMPI CLINICAL SCALES
(N = 219 men)

Scales	Scaled vectors				
	I	II	III	IV	V
L	−.059	.294	.126	.089	−.104
F	−.011	−.079	−.677	−.612	−.137
K	−.470	−.251	.060	−.046	−.317
Hs	−.257	−.062	.032	−.410	.234
D	−.094	.257	−.235	.103	.025
Hy	.516	−.025	−.375	.240	−.050
Pd	.271	−.342	.084	−.176	−.247
Mf	−.169	−.075	.109	.123	.181
Pa	.007	.036	−.304	.349	.474
Pt	.465	.387	.406	.195	−.397
Sc	−.632	.413	.531	.264	−.406
Ma	.041	−.113	−.121	−.162	.268
Si	−.001	−.415	.042	−.290	.064

Latent roots	Trace (percent)
$\lambda1 = .2095$	39.48
$\lambda2 = .1591$	29.99
$\lambda3 = .0886$	16.70
$\lambda4 = .0502$	9.46
$\lambda5 = .0232$	4.37

Trace = .5305; $\Lambda = .6098$; $F^{65}_{984} = 1.62$, $p < .01$.

employed for scale independence (in the correlational sense) during item analysis was quite minimal and the number of separate substantive dimensions involved in this set of scales is certainly less than 13. It is of interest, therefore, to examine the nature of the factor structure underlying the content scales and to do so with reference to the manner in which this structure is manifest in diverse populations. The samples selected for such analysis were: (*a*) 261 Air Force enlisted men, (*b*) 258 male psychiatric inpatients, and (*c*) 100 University of Illinois male students. Although sex (male) and geographical locale (Illinois) are common for these samples, they are assumed to vary on a large number of other demographic characteristics.

Matrices of intercorrelations among the 13 content scales were factored by the method of principal components. Factors were retained whose latent roots were greater than one. Three factors met this criterion in each of the samples.[10] The retained factors accounted for 69%, 71%, and 62% of the total scale variance in the Air Force, psychiatric, and college samples, respectively. The factor matrices were rotated to a varimax criterion. The rotated factor matrices for each of the three samples are presented in Table 18.

Factor I. The first factor in the Air Force sample is a large (53% of common variance) and general dimension of self-reported maladjustment which is substantially loaded by all but three of the MMPI content scales. Organic Symptoms, Phobias, Poor Health, and Depression are especially highly loaded on this factor in the psychi-

[10] Although the employment of Guttman's weaker, lower bound as a criterion for determining the number of factors will not be given a general defense here, it is of interest to note that in this particular instance other criteria yield equivalent results. As part of a larger methodological study, Linn (1965) analyzed the present Air Force data and found the same number of factors to be indicated by: (*a*) inflection in the curve of latent roots and (*b*) the mean square ratio between factor loadings based on the original data matrix and factor loadings based on a data matrix augmented by randomly generated variables.

TABLE 17
DISCRIMINANT ANALYSIS OF SIX INPATIENT DIAGNOSTIC GROUPS BASED ON 13 MMPI CLINICAL SCALES
(N = 176 women)

Scales	Scaled vectors				
	I	II	III	IV	V
L	−.110	−.114	−.082	−.056	−.146
F	−.146	.267	.017	.269	.232
K	.224	.047	−.242	.181	.098
Hs	.008	−.194	.016	−.344	.107
D	−.091	−.020	.254	.316	−.080
Hy	−.031	.320	.343	−.037	.223
Pd	.183	−.206	−.457	−.192	.100
Mf	−.008	−.254	−.040	.010	.080
Pa	−.071	.228	−.190	−.265	−.405
Pt	.893	.419	−.286	−.044	−.278
Sc	−.568	−.446	−.104	.022	.280
Ma	−.071	−.008	.214	.302	.086
Si	−.021	−.023	.087	.189	−.119

Latent roots	Trace (percent)
$\lambda1 = .2425$	39.80
$\lambda2 = .1561$	25.63
$\lambda3 = .1359$	22.31
$\lambda4 = .0408$	6.70
$\lambda5 = .0339$	5.56

Trace = .6092; Λ = .5695; F^{65}_{780} = 1.49, $p < .01$.

atric sample as they are in the Air Force sample.

The maladjustment factor in the Air Force sample is highly loaded by categories of physical complaint but is also clearly a general factor of psychological complaint as well. In the psychiatric sample, the factor is less general, and hence the emphasis on physical complaint is more prominent. In the college sample this trend is reversed. Here the first factor is one which predominately emphasizes Social Maladjustment, Poor Morale, and Depression. Phobias are highly loaded on Factor I, but the categories of Organic Symptoms and Poor Health are loaded on another factor (Factor III). One of the factors underlying the relations among the content scales would thus appear to be a maladjustment or complaint factor. Its generality and relative emphasis on psychological, social and somatic symptoms would seem to vary with the population studied, however.

Factor II. The second factor in the Air Force sample is primarily loaded by Authority Conflict, Hypomania, and Manifest Hostility. More moderate loadings are contributed by Psychoticism, Depression, Family Problems, and Poor Morale. In the psychiatric sample this factor is more prominent, emerging as the first factor in the analysis, and accounting for 45% of the common variance. Again, the primary loadings are on Manifest Hostility, Hypomania, and Authority Conflict. Somewhat more substantial loadings occur on Family Conflict, Psychoticism, Poor Morale, and Depression than was the case in the Air Force sample. In the college sample this factor appears to be a slightly more general one which is distinguished by a substantial negative loading on Religious Fundamentalism. In all samples, the second factor underlying the relations among content scales is one emphasizing a cynical, distrustful, exploitive attitude toward life, hostility toward others, and restless, highstrung energy (Table 8). This aggressive orientation is accompanied by generally low morale and self-esteem and indications of coming from a home with a similar orientation (FAM). In college students this

TABLE 18

ROTATED FACTOR MATRICES OF CONTENT SCALES IN THREE MALE SAMPLES

	Chanute Air Force Base normals ($N = 261$)				Kankakee psychiatric inpatients ($N = 258$)				Illinois college men ($N = 100$)			
	I	II	III	h^2	II	I	III	h^2	I	II	III	h^2
ORG	86	19	−05	79	87	19	12	80	28	35	62	59
PHO	83	16	10	73	68	27	36	67	63	09	39	56
HEA	78	19	06	64	80	14	08	66	22	45	56	57
DEP	76	47	22	85	66	63	08	84	75	38	26	78
PSY	74	48	00	77	46	68	33	78	43	59	44	72
MOR	67	46	32	77	57	67	13	79	78	36	11	76
FEM	58	−17	−22	41	20	05	75	61	−09	−17	71	54
SOC	57	03	50	57	59	16	−08	38	83	−06	−11	71
FAM	53	46	−12	50	26	73	−12	61	44	40	11	36
AUT	01	85	−05	73	03	83	05	69	30	78	06	69
HYP	11	82	13	71	19	84	23	80	27	67	20	56
HOS	41	77	−02	76	27	86	19	85	47	69	04	71
REL	−07	−01	85	73	−05	15	82	69	21	−66	04	48
Variance (percent)	53	33	14		37	45	18		40	39	21	

orientation seems to include an element of atheism or, at least, deviation from fundamentalist religious convictions.

Factor III. In the Air Force sample the third factor is defined uniquely by Religious Fundamentalism with a secondary loading on Social Maladjustment. In the psychiatric sample, Religious Fundamentalism and Feminine Interests define the factor. Feminine Interests define the third factor in college males but Religious Fundamentalism is noticeably *not* involved, and Organic Symptoms and Health Concern emerge as scales not involved in the third factor for the Air Force or psychiatric samples. The three samples rather clearly differ with respect to the manner in which Religious Fundamentalism and Feminine Interests enter into the underlying factor structure. In the Air Force and college samples, Feminine Interests load a factor characterized by both somatic and psychological complaint. In the psychiatric sample, however, Feminine Interests load a factor primarily characterized by Religious Fundamentalism. Whereas Religious Fundamentalism defines a relatively distinct factor in the Air Force and psychiatric samples (i.e., Factor III), this category is associated with the hostility factor (Factor II) in the college sample.

Interpretation of Factorial Dimensions Underlying Content Scales

The preceding analyses suggested that the number of dimensions underlying the MMPI content scales is the same for quite different populations, but that the specific structuring of these dimensions varies with the population studied. Although the meaning of the content scales is, to some extent, self-explanatory, little attempt was made to make substantive interpretations of the three factors in the different populations. Such interpretations will require, as a minimum, the employment of marker scales which will coordinate the present findings with the extensive empirical literature of the factorial structure of the MMPI. In addition to the statistical identification of factors, it will also be necessary to relate the apparent substantive nature of the factors identified to what is known about their *extratest* correlates. Such an analysis will first require a brief consideration of the considerable literature that exists on the factorial structure of the MMPI.

With the exception of certain factorially derived inventories such as Cattell's Sixteen Personality Factor Questionnaire (Cattell & Stice, 1962), the MMPI has been subjected to more factor analytic investigations than any other test in widespread use.

Although subject to considerable interpretative controversy, a rather remarkable agreement exists as to the dimensionality of this instrument. When the intercorrelations of MMPI clinical scales are factored, two substantial factors emerge which account for the vast majority of common variance. These factors are consistently marked by Welsh's (1956) *A* and *R*, respectively, which were developed for this purpose. Depending on the investigator's tolerance for percentage of variance extracted, several additional smaller factors have been identified which appear more subject to variation, as a function of scales included and populations studied, than do the first two factors.

Early factorial studies of the MMPI tended to label the first factor "personality maladjustment" (Cook & Wherry, 1950), "psychotic maladjustment" (Cottle, 1950; Wheeler, Little, & Lehner, 1951), and "anxiety" (Eichman, 1962; Welsh, 1956). More recent studies tend to interpret both poles of the first factor and to relate it to a broader theoretical context, such as "anxiety vs. dynamic integration" (Karson & Pool, 1957, 1958), "ego-weakness vs. ego-strength" (Kassebaum, Couch, & Slater, 1959), and "general complaint vs. dynamic integration" (Gocka & Marks, 1961). In a similar fashion, interpretations of the second factor of the MMPI have changed from "overactivity and recklessness" (Cook & Wherry, 1950), "neurotic adjustment" (Cottle, 1950; Wheeler et al., 1951) and "repression" (Eichman, 1962; Welsh, 1956) to the broader category of "extraversion vs. introversion" (Gocka & Marks, 1961; Karson & Pool, 1957, 1958; Kassebaum et al., 1959). Additional factors, beyond the first two, have been variously labeled as "paranoia" (Cook & Wherry, 1950; Wheeler et al., 1951), "feminine interests" (Cook & Wherry, 1950; Cottle, 1950), "somatization" and "uncoventionality" (Eichman, 1962), and "tender-minded sensitivity" (Gocka & Marks, 1961; Kassebaum et al., 1959).

During the last decade, the foregoing substantive interpretations of MMPI factors have been seriously challenged by an argumentative and prolific group of writers devoted to the demonstration of response styles and sets in the MMPI which are alleged to vitiate or, at best, severely limit the credibility of such substantive interpretations (see Rorer, 1965). Edwards and his colleagues have steadfastly maintained that the first factor of the MMPI is best thought of as reflecting "social desirability" (Edwards, 1957, 1961, 1962; Edwards & Diers, 1962; Edwards, Diers & Walker, 1962; Edwards & Heathers, 1962; Edwards & Walker, 1961; Edwards & Walsh, 1963). Others have maintained that such stylistic tendencies as "acquiescence" (Messick & Jackson, 1961) or the tendency to answer "deviantly true" (Barnes, 1956a, 1956b; Wiggins, 1962) are involved in the first factor as well. The second factor of the MMPI has been interpreted as reflecting "acquiescence," most notably by Jackson and Messick (1958, 1961, 1962). A third stylistic factor was first identified by Edwards et al. (1962) and subsequently replicated by others (Edwards & Walsh, 1964; Liberty, Lunneborg, & Atkinson, 1964; Wiggins, 1964; Wiggins & Lovell, 1965). This factor has been referred to as a "lying" factor (Edwards et al., 1962; Liberty et al., 1964) and as a "social desirability role-playing" factor (Wiggins, 1964). Although this third factor appears to be somewhat of a "pure" response style factor which is not highly related to other sources of variance in the test, it is highly loaded by special scales which are themselves correlated with the tendency to modify answers to the test in a socially desirable direction under instructions to do so (Cofer, Chance & Judson, 1949; Boe & Kogan, 1964; Hunt, 1962; Skrzypek & Wiggins, 1966; Walker, 1962; Wiggins, 1959).

The practical relevance or even existence of response styles in the MMPI has recently been called into question (Block, 1965; McGee, 1962; Rorer & Goldberg, 1965). The most effective defense of substantive interpretations of MMPI factors has been made by Block (1965), who not only challenged stylistic interpretations on logical and statistical grounds, but provided empirical data which were, to him, demand-

ing of substantive interpretation. To demonstrate that stylistic interpretations of the first two factors of the MMPI are not *sufficient,* Block developed what he considered to be a desirability-free measure of the first factor and an acquiescence-free measure of the second factor. To the extent that these two scales mark the factors involved, one must concede that something "other" than response styles are involved. More decisive, however, was Block's appeal to the long-overdue criterion of *external* evidence of substantive dimensions being measured by the first two factors. By the method of contrasted groups, Block obtained *Q*-sort descriptions by professional psychologists of high- and low-scoring subjects on the first two factors of the MMPI. These descriptions were obtained independently of the MMPI in five diverse samples of subjects under a variety of assessment circumstances. The constellation of *Q*-sort adjectives characterizing high- and low-scoring subjects on these factors led Block to conclude that the first factor of the MMPI measures "ego-resiliency," while the second factor measures "ego-control." Space limitations prohibit documentation of the full range of closely reasoned arguments which led Block to the foreging conclusion. For present purposes, it will simply be noted that Block has convincingly demonstrated the *necessity* for substantive interpretations of these factors, regardless of their degree of contamination with other sources of variance.

For reasons dictated by the availability of original protocols, a substantive interpretation of the factors underlying the MMPI content scales will here only be attempted in college populations. An attempt was made to align the three factors found in the small sample of Illinois college men (Table 18) with reference to the principal stylistic and substantive dimensions suggested by the recent factor analytic literature of MMPI clinical scales. This was accomplished in the considerably larger samples of Stanford men and women undergraduates.

Factor analysis of the intercorrelations of MMPI content scales in the sample of Illinois undergraduate men revealed three factors (Table 18). The first factor was loaded principally by Social Maladjustment, Poor Morale, and Depression. The second factor was loaded by Authority Conflict, Manifest Hostility, and Hypomania and negatively by Religious Fundamentalism. The third factor was loaded by Feminine Interests and by the Organic Symptoms and Poor Health scales.

Additional factor analyses were performed on samples of 250 men and 250 women from Stanford University. In addition to the 13 MMPI content scales, six marker variables were included to define the traditional MMPI clinical scale space. Four of these markers are subject to stylistic interpretation, while the remaining two are not. Welsh's (1956) factor Scales *A* and *R* were included as markers of the first two factors of conventional MMPI space. Wiggins' *Sd* (Wiggins, 1959) and Cofer et al.'s (1949) *Cof* were included to investigate the possible convergence of the third content factor with the third stylistic dimension previously mentioned (Edwards et al., 1962). Block's (1965) *ER-S* and *EC-5* were included as desirability-free and acquiescence-free measures of the factors he describes as "ego-resiliency" and "ego-control," respectively. As before, the intercorrelation matrices were factored by the method of principal components, and factors with latent roots greater than unity were rotated analytically to a varimax criterion. The rotated factor matrices for samples of Stanford men and women are presented in Table 19. Factor loadings less than .33 have been omitted, and the matrices have been arranged in such a way as to facilitate comparison.

Whereas three factors were obtained from the analysis of content scales in the sample of Illinois men, five factors are seen to emerge when six marker scales are included. These five factors account for 69% and 72% of the total variance in the female and male samples, respectively. The first three factors are recognizable as the same obtained in the earlier analysis. The fourth factor is a "stylistic" factor determined by the inclusion of *Sd* and *Cof*. These stylistic

TABLE 19

ROTATED FACTOR MATRICES OF CONTENT SCALES PLUS SIX MARKER VARIABLES

	Stanford women ($N = 250$)						Stanford men ($N = 250$)					
	I	II	III	IV	V	h^2	I	III	II	IV	V	h^2
MOR	81					80	79		36			77
A	80		37			88	85		40			91
SOC	79	−36				81	62	−56				81
DEP	78		38			81	78		42			82
ER-S	−57	−42	−37			67	−69		−34			63
PHO	53				34	49	47		58			57
HOS	40	72				69	68	37				69
Cof	−40			75		79	−44			71		78
FAM	37	34				33		40	40			47
PSY	33	37	56			59	53		61			69
R		−83				74	−38	−72				78
HYP		70				66	52	58				71
EC-5		−70		40		80		−80				83
AUT		55				49	48	33	34		−35	59
ORG			82			75	34		77			71
HEA			78			74			76			69
Sd				88		78				83		77
REL				72		58				75		62
FEM					85	74					87	80
Variance (percent)	50	18	15	9	8		36	17	22	16	9	

scales have little in common with content scales other than Religious Fundamentalism, although this, in itself, is an intriguing finding. The present space is such that Feminine Interests emerges as a fifth quite specific factor, distinct from Factor II.

Factor I. *Anxiety proneness versus ego resiliency.* The first factor in both samples is clearly and unambiguously marked by Welsh's *A*, which coordinates the first factor of content scales with the first factor obtained in all studies of clinical scales to date. Although Scale *A* provides a statistical identification of the factor, it does not allow a choice between stylistic and substantive interpretations. Block's *ER-S* which does not admit of stylistic interpretation has a substantial, but not unique, loading on this factor in the present analyses. It will not be argued that the present factor is free of the contaminating influence of social desirability. However the nature of the content scales which load this factor is considered of more than passing interest. Poor Morale, Social Maladjustment, and Depression have high loadings on the first factor in both groups. The item content of these scales (Table 8) suggests an individual lacking in self-confidence, who is socially inhibited and given to feelings of guilt and apprehensiveness. An individual at the other end of the implied continuum, would be characterized by self-confidence and optimism; social ascendance and poise; and a confident, resilient approach to the future. The item content of the scales which mark this factor are so close to the independent behavior descriptions obtained by Block (1965) for individuals with high and low scores on the same psychometric dimension that Block's suggested label of "ego-resiliency" is here applied to the first factorial dimension of MMPI content scales.

Factor II. *Impulsivity versus control.* The second factor is marked by Welsh's *R* and Block's *EC-5*, with *R* predominating in the sample of women and *EC-5* marking the factor in the sample of men. Although *R* is highly subject to stylistic interpretation, *EC-5* is not. In light of the recent criticisms directed at the interpretation of this dimension as "acquiescence" (Block, 1965; McGee, 1962; Rorer, 1965; Rorer & Goldberg, 1965), the burden of proof of the

utility of such an interpretation is shifted to its proponents. More germane to the present analyses are the nature of the content scales which load this factor. Hypomania has high loadings on the second factor in both samples. The items which constitute this scale emphasize exitement, restlessness, hotheadedness and overcommitment. Such items are suggestive of impulsivity or lack of control at one pole of the dimension and control, or possibly, "overcontrol" at the other. Manifest Hostility and Authority Conflict have high loadings on this factor in the female sample. The items in these scales reflect the free expression of aggressiveness and the cynical, distrustful attitude that everyone should get away with what he can. Such items are seen as consistent with the lack of impulse control suggested by the Hypomania scale. In the male sample, Manifest Hostility and Authority Conflict have smaller loadings while Social Maladjustment and Family Problems contribute more. Again, the constellation of content scales marking the second factor is highly similar to the independent behavior descriptions obtained by Block (1965) for this dimension which led him to label the factor as a "control" dimension. It is also of interest to note that Block (1965) has argued that the control dimension is expressed differently in men and women, which also appears to be the case in the present analyses.

Factor III. *Health concern.* In both samples, the third factor is characterized by high loadings on Organic Symptoms and Poor Health. The relationship between these two scales is rather obvious and would seem to warrant the general label of "health concern" for the factor. In college populations, at least, there appears to be an underlying factor of concern with health that includes both the headaches, dizziness, etc. from the Organic Symptoms content scale and the gastrointestinal complaints from the Poor Health content scale. Lacking the independent behavior descriptions, which were available for the first two factors, and lacking a factor marker which would relate the present dimension to previous ones, it seems best to view this factor as one involving "reported poor health." Considering the large number of items in the MMPI pool which relate to health, it is not surprising that such a factor would emerge. A "Poor Physical Health" factor has emerged from factor analysis of items from several of the standard MMPI clinical scales (Comrey, 1957a, 1957b, 1957c, 1958c; Comrey & Marggraff, 1958). Factor analysis of groups of MMPI scales has also yielded a "somatization" factor (Eichman, 1961; Fisher, 1964).

Factor IV. *Social desirability role playing.* This factor was rather clearly determined by the inclusion of the stylistic role-playing scales which define it: Wiggins' (1959) *Sd* and Cofer et al.'s (1949) *Cof*. As previously indicated, a considerable number of studies attest to the behavioral correlates of these scales; namely the tendency to modify answers to the MMPI in a socially desirable direction when instructed to do so. The content scale of Religious Fundamentalism has a high and unique loading on this factor in both samples. The most conservative interpretation of this finding would be that the items in the Religious Fundamentalism scale are those which are most subject to change under conditions which encourage faking. However, in view of the fact that the Marlowe-Crowne Social Desirability Scale (Crowne & Marlowe, 1960) is known to load this factor (Edwards et al., 1962; Edwards & Walsh, 1964; Liberty et al., 1964), a further inference seems justified. On the basis of the extensive documentation of the correlates of the Marlowe-Crowne scale (Crowne & Marlowe, 1964) which is known to load this factor, it seems likely that, in these college samples, individuals who describe themselves as religious, churchgoing people may be operating under a strong motive to gain social approval (Crowne & Marlowe, 1964). Such a phenomenon may be quite specific to these particular samples, however.

Factor V. *Feminine interests.* When the present set of marker scales are included in the factor analysis of MMPI content scales, Feminine Interests emerges as a specific factor. In the earlier factor analyses of content scales in several male samples

(Table 18), the Feminine Interests scale was seen to vary from sample to sample in its factorial contribution. Such a factor is reminiscent of the "feminine interests" factor which has been reported from time to time in the literature (Cook & Wherry, 1950; Cottle, 1950; Kassebaum et al., 1959; Wheeler et al., 1951) and which has exhibited considerable fluctuation from sample to sample. Interpretation of this factor must be restricted to the content of the items in the Feminine Interests scale (Table 8) since its non-MMPI correlates have not been investigated.

In summary, when the factorial dimensions of the MMPI content scales were aligned with previously reported dimensions of MMPI clinical scales, considerable convergence was evident. In a college population, the first two factors were clearly marked by Welsh's (1956) *A* and *R* which permitted their identification as the first two factors of previously reported studies. Although the possible contaminating effect of "social desirability" could not be ruled out, the first factor was interpreted as reflecting "anxiety-proneness versus ego resiliency." The second factor was interpreted as "impulsivity versus control" with less concern for the possible alternative interpretation of "acquiescence." The third factor appeared to reflect "health concern" as judged from the item content of the scales which loaded it. A relatively specific factor of "feminine interests" was identified, although its generality across populations was questioned. The possibility that high scores on the Religious Fundamentalism content scale may be associated with high approval motivation (Crowne & Marlowe, 1964) was also raised.

Implications for Clinical Interpretation of the MMPI

In several of the analyses previously presented, the MMPI clinical scales were used as a baseline or frame of reference for comparison with the content scales. These comparisons involved such issues as internal consistency, group differences, and differential diagnosis. Although the clinical and content scales were found to be similar

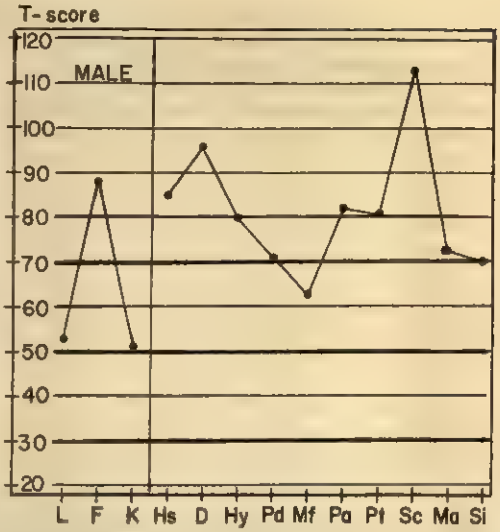

Fig. 1. Profile of hypothetical patient on standard MMPI scales.

in many respects, they should not be viewed as equivalent from the standpoint of clinical application. This point is best illustrated by use of an artificially constructed profile of MMPI clinical scale scores. Such a profile is presented in Figure 1.

The profile in Figure 1 bears a resemblance to that of "John Doe" presented by Shneidman (1951, p. 221). Shneidman's patient was a 25-year old single male whose primary diagnosis was that of anxiety reaction but for whom there were indications (especially in the MMPI) of an incipient schizophrenic reaction. The hypothetical profile in Figure 1 differs from John Doe's principally in *Pt* which is lower and in *Mf* and *Sc* (also lower) plus a slightly higher *F*. Configurally, the hypothetical profile resembles that of a 21-year old single male reported in the MMPI Atlas (Hathaway & Meehl, 1951, p. 605). The diagnosis for this patient was reactive depression with a lingering doubt concerning organic pathology. The overall elevation of the profile reported in the Atlas was much less than that of the hypothetical profile. The hypothetical profile is not easily classified within Marks and Seeman's (1963) system although this is, itself, no indictment of its typicalness (Huff, 1965). All of the promising diagnostic signs employed by Goldberg

TABLE 20
RAW SCORES ON CONTENT SCALES FOR TWO HYPOTHETICAL PATIENTS WITH IDENTICAL CLINICAL SCALE PROFILES

Content scale	Patient A	Patient B	Difference
ORG	28	5	+23
PSY	8	21	−13
HEA	7	20	−13
FEM	6	17	−11
FAM	11	6	+5
DEP	15	19	−4
HYP	8	12	−4
PHO	13	9	+4
SOC	10	13	−3
HOS	9	12	−3
AUT	13	11	+2
REL	7	9	−2
MOR	12	11	+1

(1965) classify this profile as "psychotic." Perhaps the best appeal is to "clinical experience," which for many will verify the plausibility of encountering a profile such as that in Figure 1 in a hospital setting.

The profile in Figure 1 may be used to illustrate the manner in which content scales may supplement interpretation of standard MMPI profiles. Starting with two identical 566-item protocols, each of which yielded the clinical scale profile in Figure 1, content scale scores were varied under the restriction that clinical scale scores remain the same. This was done to illustrate the point that the same profile of clinical scale scores can be obtained in two protocols which differ markedly from each other in their content scale scores. Table 20 presents the raw content scale scores for two hypothetical patients (Patient A and Patient B) each of whom produces the identical MMPI clinical profile illustrated in Figure 1.

Patient A has admitted to a large number of symptoms thought to be indicative of organic pathology. Additionally, he admits having family problems and a number of fears. In contrast, Patient B admits to a large number of psychotic symptoms of a primarily paranoid nature. He is greatly concerned about his health and admits to liking an unusual number of feminine pursuits. By comparison with Patient A, Patient B is generally more deviant with respect to content categories reflecting poor morale, mood instability, social maladjustment, and hostility.

The configuration of content scale scores of Patient B readily confirms the impression of psychopathology gained from an inspection of the clinical profile in Figure 1. This could be the profile of a paranoid schizophrenic with an underlying homosexual component and a body concern that is delusional in nature. Poor morale, social maladjustment, and hostility are, of course, compatible with this picture.

Although Patient A's raw content scale scores are sufficiently deviant to be considered those of a hospitalized patient, they are in sharp contrast to those of Patient B. By comparison, Patient A is almost exclusively concerned with organic symptoms and, to a lesser extent, family problems. Evidence of delusional thinking, health concern, feminine interests, and general maladjustment is comparatively weak for Patient A. The clinical scale profile in Figure 1 may now be viewed in a quite different light.

The present example does *not* imply that the long-awaited method of differentiating schizophrenic from brain disorders has been discovered to reside within the MMPI. It is meant to imply that a given clinical scale profile may be viewed in quite different perspectives as a function of variation in the underlying content components which determine profile elevation. The interpretative significance of content scale configurations cannot be taken at face value, and a great deal more research and experience with these scales must precede any recommendations for clinical application. Curiosity concerning the nature of patients' communications to us would seem to be a healthy interest, however, and some may prefer to adopt the MMPI content scales as the most promising procedure for satisfying this interest. It is hoped that such interim applications would be strictly supplemental to the tried and, occasionally, true procedures for clinical scale interpretation.

DISCUSSION

To encourge further investigation of the empirical properties of the content scales is to imply that they possess advantages

over the currently employed clinical scales. Since such a position is taken by the present investigator, it seems appropriate to review these claimed advantages and to discuss their relevance for both clinical and research application of the MMPI.

Viewed from the convenient hindsight of 25 years, the MMPI appears to have been poorly conceived for the purposes it was eventually to serve. The Kraepelinian categories to which it was committed were soon to pass into disfavor. Moreover, the predictive success of the individual scales in making such psychiatric categorizations was considerably less than had been anticipated. Under the impetus of an unprecedented amount of research, there was a shift of emphasis from the psychiatric to the "personological" implications of the clinical scales and the application of the scales was extended far beyond the original context of personnel decisions.

The MMPI clinical scales are poorly equipped to serve as personality trait scales for several reasons. Several of the scales lack the internal consistency which is usually taken as evidence of an organized pattern of behavior. Also, an interpretative ambiguity exists with respect to the meaning and significance of low scores on the scales since "normal" subjects rarely achieve a score of zero (Wiggins, 1962, pp. 226–227). Indeed, the hodgepodge of content which contributes to a high score on a given clinical scale is not suggestive of any consistent personality trait or structure. The fact that this makes the inventory difficult to fake would seem, at best, a mixed blessing. Given the substantive heterogeneity of the clinical scales, a configural "pattern" may be achieved in a wide variety of ways, and it seems cavalier to apply standard "blind" interpretations to such patterns, as is done in clinical practice. Finally, a minor but irritating characteristic of the scales is the extensive degree of item overlap which exists among them (Adams & Horn, 1965; Shure & Rogers, 1965).

It seems likely that the MMPI item pool, which was once considered so rich and untapped, may be too limited as a source of items for building general-purpose personality scales (Wiggins & Goldberg, 1965). This may be true with respect both to content and item characteristics, and is certainly true of the extent to which the two are confounded. Nevertheless, in the absence of any immediate replacement, it would seem unwise to abandon an inventory that has the empirical virtues, however limited, of the MMPI. Rather, it would seem appropriate to explore the utility of supplemental measures which are not encumbered by all the substantive and psychometric shortcomings of the clinical scales.

The MMPI content scales possess a respectable degree of internal consistency. This internal consistency must, in part, be attributed to homogeneous organizations of psychological, physical, and social complaints which seem appropriately combined by a cumulative scoring model (Loevinger, 1957, pp. 664–666). Although no claim is made for scale unidimensionality or Guttman-type item properties, each scale has a compelling, though prosaic, feature. Subjects who achieve high scores on the scales do so by admitting to, or claiming, an unusual amount of the substantive dimension involved. Subjects who achieve low scores claim a small amount and, by so doing, may or may not be similar to certain abnormal groups. But subjects who say they are hostile are saying just that and not that they have organic symptoms or strong religious convictions. A return to this type of Woodworthian simplicity has been long overdue.

The present study was able to provide only very limited evidence bearing on the effectiveness of the content scales in discriminating among traditional psychiatric groups. However, the preliminary evidence obtained was not discouraging in this respect. Although the burden of proof is clearly on the content scales, the superiority of scales derived by a contrasted groups strategy need not be conceded a priori when populations other than the derivation samples are involved (Hase & Goldberg, 1965).

Although apparently heterogeneous in content, covariation among content scales may be reduced to three underlying fac-

tors. The first two of these factors were found to be colinear with the first two factors consistently found in analyses of the MMPI clinical scales. This result is not surprising within the domain of MMPI items and may even reflect an upper limit on the number of parsimoniously interpretable factors within the conventionally defined questionnaire realm (Peterson, 1965). However, the content scales tend to clarify the specific manner in which the ubiquitous two factors of personality questionnaires manifest themselves within the MMPI item pool. The item content of the scales which mark these two factors lends itself readily to the substantive interpretations placed upon these dimensions by Block (1965). This is especially important when it is recognized that Block's interpretations were buttressed by independently obtained empirical evidence.

Coming from the same item pool, the content scales are no less free than the clinical scales from confounding item characteristics that lend themselves to stylistic interpretations. However, the tenor of recent critical thinking on this issue is such as to suggest that the burden of proof of the utility of stylistic interpretations has been shifted to the proponents of such styles. In any event, it seems clearer *what* is being confounded by item characteristics in the case of the content scales. Future studies of item characteristics would do well to examine their effects on substantive dimensions rather than on the poorly understood dimensions yielded by the scale construction strategy of contrasted groups. Such research would naturally be facilitated by scales composed of nonoverlapping items.

The case for further investigation of substantive aspects of the MMPI may best be presented by calling attention to a basic feature of assessment situations which has tended to be ignored or belittled by sophistic arguments. Regardless of psychologists' views of a test response, the respondent tends to view the testing situation as an opportunity for *communication* between himself and the tester or institution he represents (Leary, 1957). Obviously, the respondent has some control over what he chooses to communicate, and there are a variety of other factors which may enter to "distort" the message, many of them attributable to the testing media itself (Cattell, 1961; LaForge, 1963). Nevertheless, recognition of such sources of "noise" in the system should not lead us to overlook the fact that a message is still involved. The MMPI content scales may be closely attuned to this message and as such may provide a useful supplement to the standard clinical scales.

APPENDIX A

ITEM LISTINGS FOR MMPI CONTENT SCALES

SOC Social maladjustment (27 items)
True: 52, 171, 172, 180, 201, 267, 292, 304, 377, 384, 453, 455, 509
False: 57, 91, 99, 309, 371, 391, 449, 450, 479, 482, 502, 520, 521, 547

DEP Depression (33 items)
True: 41, 61, 67, 76, 94, 104, 106, 158, 202, 209, 210, 217, 259, 305, 337, 338, 339, 374, 390, 396, 413, 414, 487, 517, 518, 526, 543
False: 8, 79, 88, 207, 379, 407

FEM Feminine interests (30 items)
True: 70, 74, 77, 78, 87, 92, 126, 132, 140, 149, 203, 261, 295, 463, 538, 554, 557, 562
False: 1, 81, 219, 221, 223, 283, 300, 423, 434, 537, 552, 563

MOR Poor morale (23 items)
True: 84, 86, 138, 142, 244, 321, 357, 361, 375, 382, 389, 395, 397, 398, 411, 416, 418, 431, 531, 549, 555
False: 122, 264

REL Religious fundamentalism (12 items)
True: 58, 95, 98, 115, 206, 249, 258, 373, 483, 488, 490
False: 491

AUT Authority conflict (20 items)
True: 59, 71, 93, 116, 117, 118, 124, 250, 265, 277, 280, 298, 313, 316, 319, 406, 436, 437, 446
False: 294

PSY Psychoticism (48 items)
True: 16, 22, 24, 27, 33, 35, 40, 48, 50, 66, 73, 110, 121, 123, 127, 136, 151, 168, 184, 194, 197, 200, 232, 275, 278, 284, 291, 293, 299, 312, 317, 334, 341, 345, 348, 349, 350, 364, 400, 420, 433, 448, 476, 511, 551
False: 198, 347, 464

ORG Organic symptoms (36 items)
True: 23, 44, 108, 114, 156, 159, 161, 186, 189, 251, 273, 332, 335, 541, 560
False: 46, 68, 103, 119, 154, 174, 175, 178, 185, 187, 188, 190, 192, 243, 274, 281, 330, 405, 496, 508, 540

FAM Family problems (16 items)
True: 21, 212, 216, 224, 226, 239, 245, 325, 327, 421, 516
False: 65, 96, 137, 220, 527

HOS Manifest hostility (27 items)
True: 28, 39, 80, 89, 109, 129, 139, 145, 162, 218, 269, 282, 336, 355, 363, 368, 393, 410, 417, 426, 438, 447, 452, 468, 469, 495, 536

PHO Phobias (27 items)
True: 166, 182, 351, 352, 360, 365, 385, 388, 392, 473, 480, 492, 494, 499, 525, 553
False: 128, 131, 169, 176, 287, 353, 367, 401, 412, 522, 539

HYP Hypomania (25 items)
True: 13, 134, 146, 181, 196, 228, 234, 238, 248, 266, 268, 272, 296, 340, 342, 372, 381, 386, 409, 439, 445, 465, 500, 505, 506

HEA Poor health (28 items)
True: 10, 14, 29, 34, 72, 125, 279, 424, 519, 544
False: 2, 18, 36, 51, 55, 63, 130, 153, 155, 163, 193, 214, 230, 462, 474, 486, 533, 542

APPENDIX B

TABLE B1

SOCIAL MALADJUSTMENT

$\bar{X}$	σ	Group	N
16.71	(5.97)	Personality pattern disturbance (OP)	17
13.48	(8.50)	Brain disorders	23
12.80	(8.33)	Psychoneurotic disorders (OP)	15
11.92	(7.12)	Personality trait disturbance (OP)	36
9.83	(7.65)	Special symptom reaction (OP)	6
9.83	(5.41)	Psychoneurotic disorders	13
9.75	(5.42)	Transient situational disturbance (OP)	8
9.57	(5.42)	Air Force normals	261
9.55	(4.80)	Schizophrenic psychoses	85
9.33	(5.93)	Sociopathic personality disturbance	46
8.35	(5.87)	Affective psychoses	20
8.21	(6.75)	Sociopathic personality disturbance (OP)	19
8.16	(5.50)	College normals	96
8.12	(4.87)	Personality trait disturbance	17
6.33	(5.14)	Personality pattern disturbance	15
5.75	(4.04)	Brain disorders (OP)	16

TABLE B2

DEPRESSION

$\bar{X}$	σ	Group	N
16.12	(4.94)	Personality pattern disturbance (OP)	17
14.17	(9.56)	Personality trait disturbance (OP)	36
13.65	(8.92)	Brain disorders	23
12.58	(8.01)	Psychoneurotic disorders	15
12.33	(7.28)	Special symptom reaction (OP)	6
11.16	(6.99)	Sociopathic personality disturbance (OP)	19
10.73	(9.02)	Psychoneurotic disorders (OP)	15
10.11	(7.01)	Sociopathic personality disturbance	46
9.48	(6.39)	Air Force normals	261
9.00	(5.78)	Personality trait disturbance	17
8.96	(6.50)	Schizophrenic psychoses	85
8.70	(5.36)	Affective psychoses	20
6.20	(5.40)	Personality pattern disturbance	15
5.42	(4.72)	College normals	96
5.31	(5.16)	Brain disorders (OP)	16
4.13	(3.18)	Transient situational disturbance (OP)	8

TABLE B3

FEMININITY

$\bar{X}$	σ	Group	N
12.39	(4.35)	Brain disorders	23
11.81	(3.83)	Schizophrenic psychoses	85
11.55	(3.65)	Affective psychoses	20
11.05	(5.21)	Sociopathic personality disturbance (OP)	19
11.04	(3.66)	Sociopathic personality disturbance	46
10.80	(3.91)	Personality pattern disturbance	15
9.98	(3.72)	Air Force normals	261
9.77	(4.25)	Psychoneurotic disorders	13
9.67	(2.16)	Special symptom reaction (OP)	6
9.65	(3.06)	Personality trait disturbance	17
9.16	(3.42)	College normals	96
8.33	(3.35)	Psychoneurotic disorders (OP)	15
8.18	(3.41)	Personality pattern disturbance (OP)	17
8.17	(4.49)	Personality trait disturbance (OP)	36
7.94	(2.11)	Brain disorders (OP)	16
7.50	(2.20)	Transient situational disturbance (OP)	8

TABLE B4

POOR MORALE

$\bar{X}$	σ	Group	N
14.06	(5.48)	Personality pattern disturbance (OP)	17
11.61	(6.85)	Personality trait disturbance (OP)	36
11.30	(6.74)	Brain disorders	23
10.22	(4.99)	Sociopathic personality disturbance (OP)	19
10.00	(7.80)	Special symptom reaction (OP)	6
9.93	(6.82)	Sociopathic personality disturbance	46
9.80	(6.96)	Psychoneurotic disorders (OP)	15
9.00	(6.34)	Affective psychoses	20
8.92	(6.75)	Psychoneurotic disorders	13
8.31	(5.31)	Air Force normals	261
8.24	(5.41)	Schizophrenic psychoses	85
7.41	(4.73)	Personality trait disturbance	17
6.25	(5.77)	Brain disorders (OP)	16
5.13	(3.64)	Transient situational disturbance (OP)	8
4.96	(4.29)	College normals	96
3.93	(3.45)	Personality pattern disturbance	15

TABLE B5

RELIGION

$\bar{X}$	σ	Group	N
7.33	(2.58)	Special symptom reaction (OP)	6
7.16	(2.56)	Air Force normals	261
7.13	(2.36)	Brain disorders (OP)	16
6.90	(2.53)	Affective psychoses	20
6.87	(3.09)	Brain disorders	23
6.64	(2.69)	Sociopathic personality disturbance	46
6.57	(3.00)	Schizophrenic psychoses	85
6.08	(2.71)	Psychoneurotic disorders	13
5.91	(3.50)	College normals	96
5.88	(2.53)	Transient situational disturbance (OP)	8
5.83	(3.39)	Personality trait disturbance (OP)	36
5.67	(2.89)	Psychoneurotic disorders (OP)	15
5.40	(2.13)	Personality pattern disturbance	15
5.35	(2.23)	Personality trait disturbance	17
5.24	(2.17)	Personality pattern disturbance (OP)	17
5.11	(2.58)	Sociopathic personality disturbance (OP)	19

TABLE B6

AUTHORITY CONFLICT

$\bar{X}$	σ	Group	N
12.71	(4.34)	Sociopathic personality disturbance	46
11.74	(5.06)	Brain disorders	23
11.63	(4.00)	Sociopathic personality disturbance (OP)	19
11.59	(4.37)	Personality pattern disturbance (OP)	17
11.42	(4.77)	Personality trait disturbance (OP)	36
11.05	(3.61)	Air Force normals	261
10.90	(4.28)	Affective psychoses	20
10.07	(4.56)	Psychoneurotic disorders (OP)	15
9.87	(4.90)	Schizophrenic psychoses	85
9.85	(4.86)	Psychoneurotic disorders	13
9.24	(4.44)	Personality trait disturbance	17
8.81	(4.48)	Brain disorders (OP)	16
8.24	(4.18)	College normals	96
8.17	(4.36)	Special symptom reaction (OP)	6
8.00	(4.04)	Personality pattern disturbance (OP)	17
6.38	(4.14)	Transient situational disturbance (OP)	8

TABLE B7

PSYCHOTICISM

$\bar{X}$	σ	Group	N
14.43	(4.10)	Brain disorders	23
13.18	(7.26)	Personality pattern disturbance (OP)	17
12.47	(9.45)	Personality trait disturbance (OP)	36
11.38	(7.39)	Air Force normals	261
10.87	(8.82)	Sociopathic personality disturbance	46
10.33	(9.20)	Schizophrenic psychoses	85
10.25	(7.65)	Affective psychoses	20
8.63	(6.16)	Sociopathic personality disturbance (OP)	19
8.46	(6.33)	Psychoneurotic disorders	13
7.73	(5.39)	Psychoneurotic disorders (OP)	15
7.40	(7.92)	Personality pattern disturbance	15
6.50	(3.67)	Special symptom reaction (OP)	6
6.13	(5.41)	Brain disorders (OP)	16
6.12	(4.06)	Personality trait disturbance	17
5.68	(4.17)	College normals	96
4.63	(3.25)	Transient situational disturbance (OP)	8

TABLE B8

ORGANIC SYMPTOMS

$\bar{X}$	σ	Group	N
13.65	(5.98)	Personality pattern disturbance (OP)	17
12.03	(8.15)	Personality trait disturbance (OP)	36
11.17	(8.34)	Brain disorders	23
10.47	(7.80)	Psychoneurotic disorders (OP)	15
10.00	(4.90)	Special symptom reaction (OP)	6
8.08	(7.03)	Psychoneurotic disorders	13
7.31	(5.64)	Sociopathic personality disturbance	46
7.00	(5.85)	Air Force normals	261
6.81	(5.39)	Schizophrenic psychoses	85
6.47	(4.69)	Personality trait disturbance	17
6.40	(4.20)	Affective psychoses	20
5.68	(3.84)	Sociopathic personality disturbance (OP)	19
4.87	(5.08)	Personality pattern disturbance	15
3.69	(3.79)	Brain disorders (OP)	16
3.63	(3.20)	Transient situational disturbance (OP)	8
3.02	(3.23)	College normals	96

TABLE B9

FAMILY PROBLEMS

$\bar{X}$	σ	Group	N
6.56	(4.17)	Personality trait disturbance (OP)	36
6.42	(3.81)	Sociopathic personality disturbance (OP)	19
5.94	(2.82)	Personality pattern disturbance (OP)	17
5.74	(3.72)	Brain disorders	23
5.67	(3.22)	Sociopathic personality disorder	46
5.38	(3.91)	Psychoneurotic disorders	13
5.20	(3.34)	Schizophrenic psychoses	85
5.13	(3.29)	Personality pattern disturbance	15
5.06	(2.84)	Personality trait disturbance	17
4.95	(2.99)	Air Force normals	261
4.95	(2.54)	Affective psychoses	20
4.53	(3.52)	Psychoneurotic disorders (OP)	15
4.38	(3.88)	Brain disorders (OP)	16
4.11	(2.70)	College normals	96
3.83	(3.92)	Special symptom reaction (OP)	6
2.88	(1.96)	Transient situational disturbance (OP)	8

TABLE B10
HOSTILITY

$\bar{X}$	σ	Group	N
12.71	(5.07)	Personality pattern disturbance (OP)	17
11.81	(5.83)	Personality trait disturbance (OP)	36
11.24	(4.69)	Air Force normals	261
10.80	(5.63)	Affective psychoses	20
10.17	(7.63)	Brain disorders	23
10.05	(5.31)	Sociopathic personality disturbance (OP)	19
9.98	(5.59)	Sociopathic personality disturbance	46
9.60	(6.03)	Psychoneurotic disorders (OP)	15
9.00	(4.85)	College normals	96
8.87	(5.96)	Schizophrenic psychoses	85
8.46	(4.89)	Psychoneurotic disorders	13
8.33	(3.98)	Special symptom reaction (OP)	6
8.29	(4.43)	Personality trait disturbance	17
7.63	(4.31)	Transient situational disturbance (OP)	8
7.40	(5.58)	Personality pattern disturbance	15
7.00	(4.50)	Brain disorders (OP)	16

TABLE B11
PHOBIAS

$\bar{X}$	σ	Group	N
8.82	(4.19)	Personality pattern disturbance (OP)	17
8.67	(4.41)	Special symptom reaction (OP)	6
8.30	(4.93)	Brain disorders	23
7.91	(4.45)	Sociopathic personality disturbance	46
7.53	(4.53)	Psychoneurotic disorders (OP)	15
7.48	(4.14)	Schizophrenic psychoses	85
7.11	(5.02)	Personality trait disturbance (OP)	36
6.87	(4.29)	Air Force normals	261
6.80	(4.25)	Affective psychoses	20
6.58	(3.98)	Sociopathic personality disturbance (OP)	19
6.00	(4.74)	Psychoneurotic disorders	13
6.00	(3.02)	Personality trait disturbance	17
5.44	(4.08)	Brain disorders (OP)	16
5.00	(4.41)	Personality pattern disturbance	15
4.15	(2.80)	College normals	96
3.25	(2.76)	Transient situational disturbance (OP)	8

TABLE B12

HYPOMANIA

$\bar{X}$	σ	Group	N
14.94	(4.46)	Personality pattern disturbance (OP)	17
14.00	(4.89)	Sociopathic personality disturbance (OP)	19
13.80	(5.05)	Affective Psychoses	20
13.71	(5.17)	Sociopathic personality disturbance	46
13.61	(6.19)	Brain disorders	23
13.29	(3.88)	Air Force normals	261
12.89	(3.39)	Personality trait disturbance (OP)	36
12.67	(2.50)	Special symptom reaction (OP)	6
12.07	(4.51)	Psychoneurotic disorders (OP)	15
11.85	(5.55)	Psychoneurotic disorders	13
11.74	(3.77)	College normals	96
11.63	(5.67)	Schizophrenic psychoses	85
11.63	(3.34)	Brain disorders (OP)	16
11.41	(4.35)	Personality trait disturbance	17
10.38	(4.63)	Transient situational disturbance (OP)	8
9.87	(4.47)	Personality pattern disturbance	15

TABLE B13

POOR HEALTH

$\bar{X}$	σ	Group	N
9.06	(4.53)	Personality pattern disturbance (OP)	17
8.93	(5.97)	Psychoneurotic disorders (OP)	15
8.25	(4.64)	Personality trait disturbance (OP)	36
8.22	(5.18)	Brain disorders	23
8.17	(3.71)	Special symptom reaction (OP)	6
7.46	(5.47)	Psychoneurotic disorders	13
6.90	(3.26)	Affective psychoses	20
6.71	(4.10)	Personality trait disturbance	17
6.50	(4.69)	Schizophrenic psychoses	85
6.47	(4.40)	Sociopathic personality disturbance	46
6.40	(4.05)	Air Force normals	261
5.42	(3.92)	Sociopathic personality disturbance (OP)	19
4.67	(3.61)	Personality pattern disturbance	15
4.00	(2.83)	Transient situational disturbance (OP)	8
3.56	(2.68)	Brain disorders (OP)	16
3.17	(2.34)	College normals	96

REFERENCES

ADAMS, D. K., & HORN, J. L. Nonoverlapping keys for the MMPI scales. *Journal of Consulting Psychology,* 1965, **29,** 284.

ADORNO, T. W., FRENKEL-BRUNSWIK, E., LEVINSON, D. J., & SANFORD, R. N. *The authoritarian personality.* New York: Harper, 1950.

AMERICAN PSYCHIATRIC ASSOCIATION, *Diagnostic and statistical manual: Mental disorders.* Washington: APA Mental Hospital Service, 1952.

AMERICAN PSYCHOLOGICAL ASSOCIATION, Technical recommendations for psychological tests and diagnostic techniques. *Psychological Bulletin Supplement,* 1954, **51,** No. 2, Pt. 2.

AMERICAN PSYCHOLOGICAL ASSOCIATION. *American Psychologist,* (Entire issue), 1965, **20,** 857–1002.

BARNES, E. H. Factors, response bias and the MMPI. *Journal of Consulting Psychology,* 1956, **20,** 419–421. (a)

BARNES, E. H. Response bias and the MMPI. *Journal of Consulting Psychology,* 1956, **20,** 371–374. (b)

BERG, I. A. Response bias and personality: The deviation hypotheses. *Journal of Psychology,* 1955, **40,** 61–72.

BERG, I. A. The unimportance of test item content. In B. M. Bass & I. A. Berg (Eds.), *Objective approaches to personality assessment.* New York: Van Nostrand, 1959. Pp. 83–99.

BERG, I. A. Measuring deviant behavior by means of deviant response sets. In I. A. Berg & B. M. Bass (Eds.), *Conformity and deviation.* New York: Harper, 1961. Pp. 328–379.

BLOCK, J. *The challenge of response sets.* New York: Appleton-Century-Crofts, 1965.

BOE, E. E., & KOGAN, W. S. Effect of social desirability instructions on several MMPI measures of social desirability. *Journal of Consulting Psychology,* 1964, **28,** 248–251.

BRAATEN, D. Kooky personality test. *The Washington Star,* June 8, 1965.

CATTELL, R. B. Theory of situational, instrument, second order, and refraction factors in personality structure research. *Psychological Bulletin,* 1961, **58,** 160–174.

CATTELL, R. B., & STICE, G. F. *The sixteen personality factor questionnaire.* (3rd. ed.) Champaign, Ill.: Institute for Personality and Ability Testing, 1962.

COFER, C. N., CHANCE, J., & JUDSON, A. J. A study of malingering on the MMPI. *Journal of Psychology,* 1949, **27,** 491–499.

COMREY, A. L. A factor analysis of items on the MMPI depression scale. *Educational and Psychological Measurement,* 1957, **17,** 578–585. (a)

COMREY, A. L. A factor analysis of items on the MMPI hypochondriasis scale. *Educational and Psychological Measurement,* 1957, **17,** 568–577. (b)

COMREY, A. L. A factor analysis of items on the MMPI hysteria scale. *Educational and Psychological Measurement,* 1957, **17,** 586–592. (c)

COMREY, A. L. A factor analysis of items on the MMPI hypomania scale. *Educational and Psychological Measurement,* 1958, **18,** 313–323. (a)

COMREY, A. L. A factor analysis of items on the MMPI paranoia scale. *Educational and Psychological Measurement,* 1958, **18,** 99–107. (b)

COMREY, A. L. A factor analysis of items on the MMPI psychasthenia scale. *Educational and Psychological Measurement,* 1958, **18,** 293–300. (c)

COMREY, A. L. A factor analysis of items on the MMPI psychopathic deviate scale. *Educational and Psychological Measurement,* 1958, **18,** 91–98. (d)

COMREY, A. L., & MARGGRAFF, W. M. A factor analysis of items on the MMPI schizophrenia scale. *Educational and Psychological Measurement,* 1958, **18,** 301–311.

COOK, E. B., & WHERRY, R. J. A factor analysis of MMPI and aptitude test data. *Journal of Applied Psychology,* 1950, **34,** 260–266.

COOLEY, W. W., & LOHNES, P. R. *Multivariate procedures for the behavioral sciences.* New York: Wiley, 1962.

COTTLE, W. C. A factorial study of the Multiphasic, Kuder, and Bell inventories using a population of adult males. *Psychometrika,* 1950, **15,** 25–47.

CRONBACH, L. J. Coefficient alpha and the internal structure of tests. *Psychometrika,* 1951, **16,** 297–334.

CROWNE, D. P., & MARLOWE, D. A new scale of social desirability independent of psychopathology. *Journal of Consulting Psychology,* 1960, **24,** 349–354.

CROWNE, D. P., & MARLOWE, D. *The approval motive.* New York: Wiley, 1964.

DAHLSTROM, W. G., & WELSH, G. S. *An MMPI handbook: A guide to use in clinical practice and research.* Minneapolis: University of Minnesota Press, 1960.

DICKEN, C., VAN PELT, J., & BOCK, R. D. Content and acquiescence in the MMPI. UNPUBLISHED MANUSCRIPT, San Diego State College, San Diego, 1965.

DRAKE, L. E. Differential sex responses to items of the MMPI. *Journal of Applied Psychology,* 1953, **37,** 46.

EDWARDS, A. L. *The social desirability variable in personality assessment and research.* New York: Dryden Press, 1957.

EDWARDS, A. L. Social desirability or acquiescence in the MMPI? A case study with the *SD* scale. *Journal of Abnormal and Social Psychology,* 1961, **63,** 351–359.

EDWARDS, A. L. Social desirability and expected means of MMPI scales. *Educational and Psychological Measurement,* 1962, **22,** 71–76.

EDWARDS, A. L., & DIERS, C. J. Social desirability and the factorial interpretation of the MMPI. *Educational and Psychological Measurement,* 1962, **22,** 501–509.

EDWARDS, A. L., DIERS, C. J., & WALKER, J. N. Response sets and factor loadings on sixty-one personality scales. *Journal of Applied Psychology.* 1962, **46,** 220–225.

EDWARDS, A. L., & HEATHERS, L. B. The first factor of the MMPI: Social desirability or ego strength? *Journal of Consulting Psychology,* 1962, **26,** 99–100.

EDWARDS, A. L., & WALKER, J. N. Social desirability and agreement response set. *Journal of Abnormal and Social Psychology,* 1961, **62,** 180–183.

EDWARDS, A. L., & WALSH, J. A. The relationship between the intensity of the social desirability keying of a scale and the correlation of the scale with Edwards' *SD* scale and the first factor loading of the scale. *Journal of Clinical Psychology,* 1963, **19,** 200–203.

EDWARDS, A. L., & WALSH, J. A. Response sets in standard and experimental personality scales. *American Educational Research Journal,* 1964, **1,** 52–61.

EICHMAN, W. J. Replicated factors on the MMPI with female NP patients. *Journal of Consulting Psychology,* 1961, **25,** 55–60.

EICHMAN, W. J. Factored scales for the MMPI. *Journal of Clinical Psychology Monograph Supplement,* 1962, No. 15.

FISHER, J. Some MMPI dimensions of physical and psychological illness. *Journal of Clinical Psychology,* 1964, **20,** 369–375.

GILLILAND, A. R., & COLGIN, R. Norms, reliability, and forms of the MMPI. *Journal of Consulting Psychology,* 1951, **15,** 435–438.

GOCKA, E. F., & MARKS, J. B. Second-order factors in the 16 PF and MMPI inventory. *Journal of Clinical Psychology,* 1961, **17,** 32–35.

GOLDBERG, L. R. Diagnosticians vs. diagnostic signs: The diagnosis of psychosis vs. neurosis from the MMPI. *Psychological Monographs*, 1965, **79** (9, Whole No. 602).

GORDON, J. E. A communication: Snooping and testing. *The New Republic*, January 9, 1965, 28–30.

GOUGH, H. G. Theory and measurement of socialization. *Journal of Consulting Psychology*, 1960, **24**, 23–30.

HARMAN, H. *Modern factor analysis*. Chicago: University of Chicago Press, 1960.

HARRIS, R. E., & LINGOES, J. C. *Subscales for the MMPI: An aid to profile interpretation*. Unpublished manuscript, The Langley Porter Neuropsychiatric Institute, 1955.

HASE, H. D., & GOLDBERG, L. R. *The comparative validity of different strategies of deriving personality inventory scales*. Unpublished paper, Oregon Research Institute, 1965.

HATHAWAY, S. R., & MCKINLEY, J. C. A multiphasic personality schedule (Minnesota): I. Construction of the schedule. *Journal of Psychology*, 1940, **10**, 249–254.

HATHAWAY, S. R., & MCKINLEY, J. C. *The Minnesota Multiphasic Personality Inventory Manual. Revised*. New York: The Psychological Corporation, 1951.

HATHAWAY, S. R., & MEEHL, P. E. *An atlas for the clinical use of the MMPI*. Minneapolis: University of Minnesota Press, 1951.

HUFF, F. W. Use of actuarial description of abnormal personality in a mental hospital. *Psychological Reports*, 1965, **17**, 224.

HUNT, D. E. *Personality patterns in adolescent boys*. Progress Report No. 7, PHS Grant M-3517, Syracuse University, 1962.

JACKSON, D. N., & MESSICK, S. Content and style in personality assessment. *Psychological Bulletin*, 1958, **55**, 243–252.

JACKSON, D. N., & MESSICK, S. Acquiescence and desirability as response determinants on the MMPI. *Educational and Psychological Measurement*, 1961, **21**, 771–790.

JACKSON, D. N., & MESSICK, S. Response styles and the assessment of psychopathology. In S. Messick and J. Ross (Eds.), *Measurement in personality and cognition*. New York: Wiley, 1962. Pp. 129–155.

KAISER, H. F. The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, 1958, **23**, 187–200.

KARSON, S., & POOL, K. B. The construct validity of the Sixteen Personality Factors Test. *Journal of Clinical Psychology*, 1957, **13**, 245–252.

KARSON, S., & POOL, K. B. Second-order factors in personality measurement. *Journal of Consulting Psychology*, 1958, **22**, 299–303.

KASSEBAUM, G. G., COUCH, A. S., & SLATER, P. E. The factorial dimensions of the MMPI. *Journal of Consulting Psychology*, 1959, **23**, 226–236.

LAFORGE, R. Interpersonal domains or interpersonal levels? A validation of Leary's "MMPI Level I indices." Paper read at the Western Psychological Association Meetings, Santa Monica, April, 1963.

LEARY, T. *Interpersonal diagnosis of personality*. New York: Ronald Press, 1957.

LIBERTY, P. G., JR., LUNNEBORG, C. E., & ATKINSON, G. C. Perceptual defense, dissimulation and response styles. *Journal of Consulting Psychology*, 1964, **28**, 529–537.

LINGOES, J. C. MMPI factors of the Harris and the Wiener Subscales. *Journal of Consulting Psychology*, 1960, **24**, 74–83.

LINN, R. H. A Monte Carle approach to the number of factors problem. Unpublished doctoral dissertation, University of Illinois, 1965.

LOEVINGER, J. Objective tests as instruments of psychological theory. *Psychological Reports Monograph*, 1957, **3**, 635–694.

LOEVINGER, J. Measuring personality patterns of women. *Genetic Psychology Monographs*, 1962, **65**, 35–136.

MCGEE, R. K. Response style as a personality variable: By what criterion? *Psychological Bulletin*, 1962, **59**, 284–295.

MARKS, P. A., & SEEMAN, W. *The actuarial description of abnormal personality*. Baltimore: Williams & Wilkins, 1963.

MEEHL, P. E. The dynamics of "structured" personality tests. *Journal of Clinical Psychology*, 1945, **1**, 296–303.

MEES, H. L. Preliminary steps in the construction of factor scales for the MMPI. Unpublished manuscript, University of Washington, 1959.

MESSICK, S., & JACKSON, D. N. Acquiescence and the factorial interpretation of the MMPI. *Psychological Bulletin*, 1961, **58**, 299–304.

NORMAN, W. T. Relative importance of test item content. *Journal of Consulting Psychology*, 1963, **27**, 166–174.

PETERSON, D. R. The scope, generality and meaning of verbally defined "personality" factors. *Psychological Review*, 1965, **72**, 48–59.

ROKEACH, M. *The open and closed mind*. New York: Basic Books, 1960.

RORER, L. G. The great response-style myth. *Psychological Bulletin*, 1965, **63**, 129–156.

RORER, L. G., & GOLDBERG, L. R. Acquiescence in the MMPI? *Educational and Psychological Measurement*, 1965, **25**, 801–817.

SCHNEIDMAN, E. S. *Thematic test analysis*. New York: Grune & Stratton, 1951.

SHURE, G. H., & ROGERS, M. S. Note of caution on the factor analysis of the MMPI. *Psychological Bulletin*, 1965, **63**, 14–18.

SKRZYPEK, G. J., & WIGGINS, J. S. Contrasted groups vs repeated measurement designs in the evaluation of social desirability scales. *Educational and Psychological Measurement*, 1966, **26**, 131–138.

STAGNER, R. The gullibility of personnel managers. *Personnel Psychology*, 1958, **11**, 347–352.

STERN, G. G., STEIN, M. I., & BLOOM, B. S. *Methods in personality assessment*. Glencoe: Free Press, 1956.

ULLMANN, L. P., & WIGGINS, J. S. Endorsement fre-

quency and the number of differentiating MMPI items to be expected by chance. *Newsletter of Research in Psychology,* 1962, **4,** 29–35.

WALKER, J. S. An examination of the role of the experimentally determined response set in evaluating Edwards' Social Desirability Scale. *Journal of Consulting Psychology,* 1962, **26,** 162–166.

WELSH, G. S. Factor dimensions *A* and *R*. In G. S. Welsh & W. G. Dahlstrom (Eds.), *Basic readings on the MMPI in psychology and medicine.* Minneapolis: University of Minnesota Press, 1956. Pp. 264–281.

WHEELER, W. M., LITTLE, K. B., & LEHNER, F. J. The internal structure of the MMPI. *Journal of Consulting Psychology,* 1951, **15,** 134–142.

WIGGINS, J. S. Interrelationships among MMPI measures of dissimulation under standard and social desirability instructions. *Journal of Consulting Psychology,* 1959, **23,** 419–427.

WIGGINS, J. S. Strategic, method, and stylistic variance in the MMPI. *Psychological Bulletin,* 1962, **59,** 224–242.

WIGGINS, J. S. Convergences among stylistic response measures from objective personality tests. *Educational and Psychological Measurement,* 1964, **24,** 551–562.

WIGGINS, J. S., & GOLDBERG, L. R. Interrelationships among MMPI item characteristics. *Educational and Psychological Measurement,* 1965, **25,** 381–397.

WIGGINS, J. S., & LOVELL, V. R. Communality and favorability as sources of method variance in the MMPI. *Educational and Psychological Measurement,* 1965, **25,** 399–412.

WIGGINS, J. S., & VOLLMAR, J. The content of the MMPI. *Journal of Clinical Psychology,* 1959, **15,** 45–47.

(Received April 29, 1966)

Vol. 80, No. 23 Whole No. 631, 1966

Psychological Monographs: General and Applied

TRANSFER OF RESPONSE IN VISUAL RECOGNITION SITUATIONS AS A FUNCTION OF FREQUENCY VARIABLES[1]

ARNOLD BINDER
University of California, Irvine

AND

W. K. ESTES
Stanford University

A series of 6 experiments investigated the principles required to account quantitatively for responses of human Ss to new combinations of cues following discrimination learning. In the training phase of each experiment, Ss learned to make identifying responses (numerical labels) to sets of stimuli (pairs of nonsense syllables) under a paired-associate procedure. After a fixed number of acquisition trials Ss were tested on stimulus compounds involving new combinations of the training cues. In all experiments a substantial proportion of variance in the test data was accounted for by a model embodying the additive rule and the probability matching rule of stimulus sampling theory. In cases when training had been conducted under standard discrimination paradigms, responses to test compounds were quite well accounted for by this model without auxiliary principles. Ss exhibited preferences for low ambiguity cues in test compounds only when this preference had been differentially reinforced during training. Under a variety of circumstances, predictions from the stimulus sampling model could be improved by the addition of a "relative novelty" principle, stating that, other things being equal, Ss tend to sample from test compounds the cues that had occurred least frequently during previous training.

In this study we are concerned with principles governing transfer of response to new combinations of cues following discrimination training under relatively simple stimulus conditions. In general, one expects response to new situations to be determined jointly by stimulus characteristics of the given situation and by the relevant learning history. To reduce an overwhelmingly complex problem to manageable proportions for preliminary quantitative analysis, the authors and their associates have pursued the strategy of limiting consideration to situations involving sets of simple, discrete, clearly discriminable cues,[2] while studying transfer as a function of frequencies of relevant events during previous learning experiences.

From a rather extensive series of studies conducted within this context, two principles of some generality have emerged. The first of these is the *additive rule*, which has been taken as a basic response axiom in stimulus sampling models for learning (e.g., Atkinson & Estes, 1963; Estes, 1959b). If each member of a population of cues has been uniformly associated with some one member of a set of alternative responses during training, then according to the additive rule the probability of a response on a test trial is equal to the proportion of the cues present on the test which have been associated with the given response. For example, in an experiment reported by one of the authors (Atkinson & Estes, 1963, p.

[1] This research was supported by Research Grants MH 02170 and MH 11792 from the National Institute of Mental Health of the National Institutes of Health, United States Public Health Service, Contract Nonr-908(16) between the Office of Naval Research and Indiana University, and Contract Nonr-225(73) between the Office of Naval Research and Stanford University. Reproduction in whole or in part is permitted for any purpose of the United States Government.

[2] In accord with the previously stated position of one of the authors (Estes, 1959a, pp. 455–456), we shall take the viewpoint of the experimenter in specifying what is meant by *cues*. Thus, they are those aspects of a stimulus situation that are independently manipulated in a given experimental context. In the examples that follow, the letters *a*, *b*, *c*, etc. designate separate cues.

193), a response A_1 was reinforced during training in the presence of the set of cues *abc* and response A_2 in the presence of the set *def*; probability of response A_1 on a test trial in the presence of a new sample *abd* was predicted to be ⅔ by the additive rule (and proved to be .669 in the data).

The second principle, which may be termed the *matching rule*, has arisen from two quite different lines of investigation—studies of probability learning (e.g., Estes, 1957, 1964; Estes & Straughan, 1954) and studies of frequency effects in visual recognition situations (Binder, 1963; Binder & Feldman, 1960). According to the matching rule, if two or more different responses have been reinforced in the presence of a given cue during training, whether this cue has appeared alone or as a component of more complex patterns, then on any later test the probability that any one of these responses will be evoked by the given cue is equal to its relative frequency of reinforcement. To illustrate in terms of the Binder and Feldman (1960) study, subjects (*S*s) first learned discriminations between combinations of geometric figures which we shall denote, for convenience, by small letters, *a*, *b*, *c*, etc. Following training in which a combination of cues *ac* always had A_1 as the reinforced response and the combination *ad* always had A_2 as the reinforced response, but with the first combination occurring twice as often as the second in each training block, *S*s were tested on cue *a* alone. Predicted response frequencies for the group of *S*s, according to the matching rule, were 34.7 A_1's to 17.3 A_2's; the observed frequencies were 36 A_1's and 16 A_2's (Binder & Feldman, 1960, pp. 11–12). Numerous studies have provided quantitative support for these principles when transfer is tested following standard discrimination or paired-associate training (e.g., Binder, 1963; Binder & Feldman, 1960; Estes, Burke, Atkinson, & Frankmann, 1957; Feldman, 1963; Peterson, 1956; Schoeffler, 1954), with some recent studies adding the qualification that cues which covary during training may come to act as units, or "elements," in carrying transfer effects (Estes & Hopkins, 1961; Friedman, 1966; Friedman & Gelfand, 1964).

In the series of experiments to be reported, we have tried to provide more stringent tests of these two principles of transfer, and at the same time to assess various auxiliary rules that have been proposed, by introducing various modifications into standard training paradigms. The various auxiliary hypotheses will be spelled out in connection with the specific experiments designed to test them. To anticipate the principal results, we may indicate at the outset that the studies will provide further support for the matching and additive rules, yield only negative evidence with respect to all of the other hypotheses we had in mind at the outset of the investigation, and uncover the operation of a "relative novelty principle" which has been important in other theoretical contexts (e.g., Broadbent, 1958) but had not previously been suspected, by us at least, to be involved in the types of transfer situations under consideration.

The experiments follow a common overall plan in which *S*s first learn identifying responses to sets of stimuli by a paired-associate procedure, then after a fixed number of acquisition trials, are tested on new stimulus compounds comprising two or more of the training cues recombined in various ways. Thus every test stimulus is a compound including components of previously learned stimuli. Learning conditions and cue relationships are manipulated experimentally and relative frequencies of responses to test combinations are observed. An important aspect of our procedures is that *S*s are never prepared for the transfer tests by any special instructions. Thus results of the transfer tests may be taken as evidence regarding what was learned during discrimination training, without contamination by higher order problem-solving strategies or the like.

EXPERIMENT I

In this and the subsequent experiments, we shall take as a base-line predictions derivable from the matching rule and the additive rule, which, within the context of stimulus sampling theory, characterize the *component model* (Estes, 1959b). Systematic deviations from these predictions may

be taken to indicate effects of other factors or relationships. This first experiment was contrived as a broad spectrum exploration of several factors whose effects, if any, are confounded in standard discrimination-transfer experiments. The stimulus response relationships during training are most simply portrayed by means of the schema in Table 1, the actual cues being represented by small letters *a* through *f* and the responses by the letters *i*, *j*, and *k*.

The design may be regarded as comprising two standard discrimination tasks which are learned concurrently by *S*s. In the first of the two discriminations, *S*s must learn to make response *i* to stimulus combination *ab* and response *j* to stimulus combination *ac*, the discrimination thus involving two stimulus patterns which have an element, *a*, in common. The second discrimination involves learning to make response *i* to compound *de* and response *k* to *df*, again the discrimination involving two patterns with a common cue.

To the extent that the component model correctly represents learning in this situation, we should expect that at the end of training, cues *b*, *c*, *e*, and *f* would be associated with responses *i*, *j*, *i*, and *k*, respectively; that cue *a* would be equally likely associated with *i* and with *j*; and that cue *d* would be equally likely to be associated with *i* and with *k*. If, then, the *S* were tested on a new combination *ae*, the predicted response probabilities to the compound would be .75 for response *i* and .25 for response *j*. The basis for the prediction is as follows. On the test with *ae*, the *S* is equally likely to sample either of the two component cues. If he samples cue *e*, this necessarily leads to response *i*, but if he samples *a* it leads to responses *i* and *j* with equal probabilities; thus the total predicted probability is .25 for *j* and .25 + .50 = .75 for *i*. Predictions for other test combinations may be similarly calculated for the component model and will be presented and discussed in connection with the results of the testing phase.

We are well aware that the component model, taken by itself, cannot possibly handle the combined results of the learning and testing phases of this experiment. For,

TABLE 1
DESIGN OF EXPERIMENT I

Training: Syllable combinations	Training: Responses	Testing syllable combinations
ab	*i*	*ae*
ac	*j*	*be*
de	*i*	*ad*
df	*k*	*dc*
		bc
		cf

if *S*s responded solely in terms of associations between the separate cues and the correct response, they could never reach a criterion of 100% correct responding during discrimination training. Even at the asymptote of learning, the probability of a correct response, to, say, *ab* would be only .75 according to the same reasoning just sketched in connection with test combination *ae*. It is well known that simple discriminations like these will readily be mastered to a strict criterion of 100% proficiency by human *S*s (or, for that matter, monkeys, rats, or pigeons) and thus that any discrimination-transfer theory must involve some additional principle.

In some respects the simplest augmented theory is the *mixed model*, first preposed by Estes and Hopkins (1961) and developed quantitatively by Atkinson and Estes (1963, pp. 243–249). According to the mixed model, during discrimination training, associations are formed not only between the various individual cues, such as *a*, *b*, and *c*, and the correct response, but also between *pattern cues* and the correct response. Moreover, pattern cues, once learned, dominate the lower order component cues. Considering only the first row of Table 1, according to the mixed model, at the end of training the *S* would have formed associations between the pattern cue *ab* and response *i*, between component cue *a* and either response *i* or response *j*, and between component cue *b* and response *i*, but would be responding solely in terms of the pattern cue and thus making 100% correct responses upon presentations of the training combination *ab*. When tests on new compounds are given after discrimination training, the training and testing situations have no pattern cues

in common and thus with regard to predictions about behavior on the transfer tests, the mixed model reduces to the component model, predictions being generated as illustrated above.

In another type of augmentation of the component model, associated with the discrimination theories of Restle (1955) and Atkinson (1961), among others, it is assumed that associations between component cues and reinforced responses develop during training much as assumed in the component model but that *S* also learns to respond selectively to cues which are reliable predictors of reinforcing events. For example, *Ss* learn to select or attend to cues *b* and *c* in the first two rows of Table 1, and to ignore or "adapt to" common cues, such as cue *a* in the first two rows of Table 1, which are not uniformly correlated with reinforcement of any one response during training. If this latter type of theory is correct in essentials, one might expect the adaptation or selective perception to carry over to some extent from training to testing situations, thus leading *S* to be less likely to sample ambiguous cues than unambiguous cues when they occur in test compounds. If, for example, *S* were more likely to sample cue *e* than cue *a* in test compound *ae*, one would expect his response probabilities to deviate from those predicted by the component model, presumably in the direction of a higher probability of response *i* than allowed for by the model.

Thus our plan is first to evaluate the test data against predictions from the component model and then to use any significant deviations from component model predictions as indicators of other perceptual or response processes not taken into account in the mixed model.

METHOD

Subjects. The 80 *Ss* were Indiana University undergraduate students who participated in this experiment to fulfill a requirement of their course in introductory psychology.

Apparatus. The stimulus materials were reproduced on 2 × 2 inch slides and projected by two random access projectors onto a matte screen. The sequence of slides projected was controlled by two banks of switches, one bank was used for each block of learning trials. Thus, the experimenter was able to program for two blocks in advance. Four booths were arranged in an arc so that *Ss* who sat in them could see the stimuli on the screen equally well. The booths were lined with acoustical tile on the sides and top, and contained desks in their rear parts. The space in the rear above the desk, bounded by sides and top, was open so that the projection screen could be seen. There was a hole in the center of each desk, and a roll of Esterline-Angus paper passed directly under the hole so that *S* could write his response on the paper. Movement of the paper was coordinated with stimulus presentation by a cam timer.

Stimulus materials. Pairs of nonsense syllables were used as stimuli and numbers as responses. The basic list of syllables used to form the pairs consisted of VOP, GAK, CYQ, ZIR, TEF, and FUH. These were taken from Archer's (1960) listing and have association values between 31% and 38%. The number responses were the digits 1, 2, and 3.

Procedure. There were two phases to the experiment, an acquisition phase and a testing phase. During acquisition, *Ss* learned syllable-number combinations by anticipation with correction. Stimulus exposure was 7 seconds, response exposure 3 seconds, intertrial interval 3 seconds and delay after warning buzzer 1 second. All times were within ±5% tolerance.

Each learning block consisted of eight paired associates, and the learning phase proceeded to a fixed number of eight blocks. Since each stimulus consisted of two nonsense syllables, both the right-left and left-right order of these syllables were presented in each block. Moreover, during the first block the two orders of each syllable combination were always presented on successive trials although the order of presentation was randomized over different pairs. Beginning with the second block, the order of occurrence of each stimulus and its associated response was determined completely randomly for each even-numbered trial block. The reverse order of presentation to that of the immediately preceding block was used for each odd-numbered block.

A 1-minute rest period separated these eight paired-associate blocks and the subsequent testing phase. The *Ss* were told: "In this phase of the experiment several new syllables will be presented. Each of these new syllables will be somewhat different from the ones to which you learned to associate the numbers. Also, during this new phase of the experiment the syllables shown in the first part of the experiment will appear on the screen. When these old syllables appear, their numbers will also appear just as in the first part of the experiment; however, when new syllables appear, no number will appear in association with them." They were further told to "respond to old syllables just as you did in the first part of the experiment" and to give that response to new syllables "which you think is most closely associated."

The first part of the testing phase was simply another learning block of randomly ordered paired associates, in the same manner as the last eight

blocks. Following this block came further blocks of paired associates as before, but interspersed among the trials of these blocks were the test figures consisting of novel compounds of the nonsense syllables shown earlier. These novel compounds were exposed for 7 seconds and no numbers were shown in association with them. There were 12 such test compounds and these were interspersed randomly among the four blocks of paired associates which followed the initial block of the testing phase. The *S*s were forced to respond to all test compounds.

The use of the stimulus materials is illustrated in Table 1. The single letters in the syllable combination column refer to nonsense syllables which were assigned randomly to the letters from the pool listed above; thus *ab* might for a given group of four *S*s be FUH OAK. Numbers were assigned randomly to the letters in the response column in a similar manner, so that *i*, *j*, *k*, could take on the values 1, 2, or 3.

The learning blocks and the number of test trials were actually twice as long as the number of entries in each column since, in both the training and the testing phases, the forward (e.g., *ab*) and the reverse (*ba*) of each syllable combination were shown to each *S*.

RESULTS AND DISCUSSION

A summary of the responses of all *S*s in Experiment I to the test compounds may be seen in Table 2. The heading for the two columns of frequencies for a given response includes a designation of the stimulus combinations with which that response was associated during training. The left-hand entry under each subspanner heading, under the column head marked "Both," shows the number of responses to both left-right arrangements of the test figure given in the stub column. Thus, each of the 80 *S*s contributes two tallies over a grouping of these entries, where a grouping refers to the three entries under "Both" in a single row. For example, in Table 2, 109 of the responses to the test figures *ae* and *ea* (both given to all *S*s) were the numbers previously associated with *ab*, *ba*, *de*, and *ed* (coded as *i*, which would be 1, 2, or 3 depending upon the outcome of the randomization); similarly 36 of the responses to *ae* and *ea* were the number associated with *ac* and *ca*, and 15 the number associated with *df* and *fd*.

The right-hand entry under each subspanner heading (column headed "First") gives the number of times the indicated response was given when the test figure shown

TABLE 2

TEST RESPONSES OF ALL SUBJECTS IN EXPERIMENT I

Test figure	Response (and associated training cues)					
	i(ab-de)		*j(ac)*		*k(df)*	
	Both	First	Both	First	Both	First
ae	109	53	36	18	15	9
be	128	62	16	8	16	10
ad	76	41	50	19	34	20
dc	52	23	84	43	24	14
bc	57	28	88	45	15	7
cf	16	5	61	32	83	43

to the extreme left in the given row (or its reverse) occurred the first time in the testing phase. For this purpose, *ae* and *ea*, for example, are considered equivalent, and only the one of these which occurred first was included in this tabulation.

Since, in all phases of all experiments, each stimulus combination involved was presented to each *S* in both left-right arrangements (*ae* and *ea*, etc.), we shall avoid circumlocution throughout the remainder of this presentation by using a single label (e.g., *ae*) to represent both arrangements of a pair of cues except when a distinction between the two arrangements is specifically intended.

Since there is ample evidence that the phenomena associated with choice of response to ambiguous cues are functions of the level of previous learning achieved, separate analyses were made on the basis of grouping by error rates over asymptotic trials. Using the distribution of errors to the original syllable compounds which were made during the testing phase, *S*s were divided into two groups on the basis of number of errors so that as nearly as possible 50% were in the high group and 50% in the low group. This division led to a grouping of 41 with four or fewer errors (the numbers of *S*s with 0, 1, 2, 3, and 4 errors, respectively, being 13, 10, 9, 4 and 5) and 39 with five or more errors (range 5–27). Table 3 is like Table 2 except that the upper part of Table 3 contains only the data for *S*s with four or fewer errors and the lower part of Table 3 the data for *S*s with five or more errors.

In order to compare the pattern of test

TABLE 3

TEST RESPONSES OF HIGH- AND LOW-ERROR SUBJECTS IN EXPERIMENT I

Test figure	Response (and associated training cues)					
	i(ab-de)		*j(ac)*		*k(df)*	
	Both	First	Both	First	Both	First
Low-error *S*s						
ae	67	33	12	6	3	2
be	74	36	5	3	3	2
ad	45	23	22	10	15	8
dc	19	9	53	28	10	4
bc	28	15	50	24	4	2
cf	4	1	34	17	44	23
High-error *S*s						
ae	42	20	24	12	12	7
be	54	26	11	5	13	8
ad	31	18	28	9	19	12
dc	33	14	31	15	14	10
bc	29	13	38	21	11	5
cf	12	4	27	15	39	20

results with that predicted by the component model, we have first calculated a priori theoretical values on the assumption that test behavior was strictly a function of the combination of cues presented. Following the procedure illustrated in the introduction to this experiment, we obtained theoretical response proportions, then converted these to the theoretical frequencies exhibited in Table 4. The values in the left-hand portion of the table are to be compared with those given in the columns labeled "Both" in Table 2, and those in the right-hand portion with those given under "Both" in the upper half of Table 3.

It is apparent at a glance, firstly, that the gross patterns of observed values are reflected in the a priori predictions, and secondly, that there are some substantial quantitative disparities. Among the latter, the most annoying, from the standpoint of our present purposes, are the instances of rather large frequencies in cells for which zeros are predicted. These may be termed cases of "inappropriate responding," for the expectation that the cells in question should be empty follows, not only from the component model, but from any theory which assumes test behavior to be determined by the previous learning experiences. Further analysis shows that the greater part of the inappropriate responding is attributable to a number of high-error *S*s who evidently respond essentially at random on test trials. Thus we shall confine the remainder of this analysis to the low-error *S*s, for whom the frequencies of inappropriate responding are relatively small.

To obtain the best baseline from which to evaluate deviations from component model predictions in the case of the low-error *S*s, we must correct the predictions given in Table 4 for the observed level of inappropriate responding. In effect, we assume that on some proportion p of test trials, *S*s' behavior was determined by the cues presented and on the remaining proportion $1 - p$ by irrelevant aspects of the test situation. To estimate p, we set up the observation equations

$$79/82 = p + (1 - p)\tfrac{2}{3}$$

$$74/82 = p + (1 - p)\tfrac{1}{3}$$

and

$$78/82 = p + (1 - p)\tfrac{2}{3}$$

on the basis of the data for tests on *ae*, *be*, *bc*, and *cf* (the last two being identical), sum both sides of these equations, and solve for p, obtaining

$$\hat{p} = .861.$$

This estimate was used to compute the adjusted theoretical frequencies for the low-error *S*s shown in Table 5. Test figures *be*, *bc*, and *cf*, in which each component cue had been associated with only one response during training, and figure *ad*, in which each cue had been associated with two different

TABLE 4

A PRIORI PREDICTIONS OF TEST RESPONSES IN EXPERIMENT I FROM COMPONENT MODEL

Test figure	Response					
	All subjects			Low-error subjects		
	i	*j*	*k*	*i*	*j*	*k*
ae	120	40	0	61.5	20.5	0
be	160	0	0	82	0	0
ad	80	40	40	41	20.5	20.5
dc	40	80	40	20.5	41	20.5
bc	80	80	0	41	41	0
cf	0	80	80	0	41	41

TABLE 5

ADJUSTED COMPONENT MODEL PREDICTIONS OF TEST RESPONSES FOR LOW-ERROR SUBJECTS IN EXPERIMENT I

Test figure	Response					
	i		*j*		*k*	
	Theoretical	Observed	Theoretical	Observed	Theoretical	Observed
ae	56.7	67	21.4	12	3.9	3
be	74.4	74	3.8	5	3.8	3
ad	39.2	45	21.4	22	21.4	15
dc	21.4	19	39.2	53	21.4	10
bc	39.1	28	39.1	50	3.8	4
cf	3.8	4	39.1	34	39.1	44

responses, clearly yield no significant deviations from the component model predictions.

Test figure *bc* provides an opportunity to assess any effect of response frequency (regardless of stimulus) during training, since the response associated with *b* was reinforced twice as often on training trials as the response associated with *c*. However, there proves to be no trace of any preference for the former response on *bc* test trials, the observed deviation from equality, though probably insignificant (see below), being in the opposite direction.

For the two figures, *ae* and *dc*, in which one cue had been associated with a single response and one with two responses during training, further analysis is required. In each of these instances, the frequency of the response associated with the "unambiguous" cue (i.e., the one with a single reinforced response during training) was appreciably in excess of the component model prediction. Since each *S* contributed two responses to each test figure, a standard statistical test of these discrepancies, assuming Bernoulli trials, is not appropriate. We can, however, use such a test on the data for the "First" columns of Table 3, with the component model frequencies in Table 5 all divided by 2. The χ^2 of 2.84 computed on this basis for figure *ae* is not significant at the .05 level (and the same is true for *be*, *bc*, *cf*, and *ad*). For figure *dc*, the χ^2 of 8.06 has a probability between .01 and .02 with 2 *df*.

The fact that test responses deviate significantly from the component model patterns for *dc* but not for *ae* is of some interest. In the former case, one might characterize the deviation in terms of a preference for the less ambiguous, or more valid, cue in the test figure, either of which seems intuitively reasonable enough. But in the case of *ae*, the same preference should be operating, and one would think it should be enhanced since response *i* had been associated with both the test cues during training. However, not too much can be made of this disparity since the two χ^2 values are close to, though on opposite sides of, the boundary of the .05 critical region. Moreover, the power of χ^2 (defined exactly at the limit and approximately for fixed sample size) is a function of degrees of freedom.

We next proceed to explore somewhat different experimental paradigms, in which any tendencies toward selective sampling of test cues on the basis of ambiguity or validity might be expected to show up more clearly.

EXPERIMENT II

A new arrangement of stimulus-response relationship, summarized in Table 6, was contrived to permit any preference for more valid, or less ambiguous, cues to operate on test trials, while eliminating the possibility that some of the cues involved in

TABLE 6

DESIGN OF EXPERIMENT II

Training		Testing syllable combinations
Syllable combinations	Responses	
ab	*i*	*ad*
ac	*j*	*bd*
bc	*k*	*cd*
dx	*l*	

the transfer tests might have been selectively ignored during training. The design for the training phase included the first two stimulus-response pairs of Experiment I, together with an item in which the nonoverlapping cues of the first two combinations were paired with a new response. Clearly, *S*s could not reach perfect performance on these three items if they systematically avoided sampling any of the cues, and hence we could assume that by the end of training associations would necessarily have been formed between cue a and responses i and j; cue b and responses i and k; and cue c and responses j and k; all of these cues being equal with respect to ambiguity and validity. The fourth training combination combined a cue d with a cue, denoted by x in Table 6, which was different from one presentation of the combination to the next. The consistent component, d, necessarily associated with the response l by the end of training, was paired on test trials with each of the cues from the other three combinations. If total response frequencies proved to deviate significantly from component model predictions in the same direction observed in the case of test combinations ae and dc of Experiment I, we hoped to be able to conclude that the factor responsible was differential cue ambiguity, or validity, per se, and not a difference in the consistency with which more or less ambiguous cues had been sampled by *S*s on training trials.

METHOD

Subjects. Sixty-two *S*s, all students of introductory psychology, were used.

Apparatus. The stimuli were programmed and presented as in Experiment I, and booths were identical to those described above except for the surfaces of the desks. Instead of the hole, each desk contained a box with eight Plexiglas windows and a push-button below each of these windows. There was also a small light between each button and window to provide a signal for the *S* that his response had registered in the apparatus. The button-box assembly covered the writing hole since the Esterline-Angus paper was not used. The buttons pushed by *S*s were recorded on paper tape by a Friden punch.

The possible responses were lettered on the Plexiglas windows, behind which were a series of bulbs. The bulbs served to indicate the periods during which the *S*s could respond and were under the control of the cam-timer.

Stimulus materials. As before, pairs of nonsense syllables were used as stimuli and numbers as responses. There were two basic lists of syllables: The first, List A, consisted of VOP, GAK, CYQ, ZIR, TEF, and FUH, and the second, List B, of QAR, BYM, KEC, YIN, WOX, and CUV. (All had 31–38% association value according to Archer, 1960). The responses were the digits 1, 2, 3, 4.

Procedure. There were again acquisition and testing phases, with sequencing, timing, instructions, and method of presentation as in Experiment I. Each learning block consisted of eight paired-associate presentations, and training proceeded for eight blocks. Then six test trials were given, randomly interspersed among two blocks of training trials which followed an initial retraining block in the testing phase. A restriction was imposed upon this random interspersion such that exactly four tests occurred in the first and two in the second of the final two blocks of the testing phase. The purpose of this arrangement was to ensure that the state of learning of cue-response relationships would not suffer any appreciable retention loss over the test series.

The letter x, appearing in the bottom row of Table 6, differed in status from the other letters in that the particular nonsense syllable substituted for it varied from block to block. If FUH was randomly assigned to a, for example, FUH occurred for a in every block of training trials; but in the case of x, a group of syllables was assigned, one of these occurring in each block and all occurring during the course of learning.

The syllables from List A were randomly assigned to the letters, a, b, c, d, and the syllables from List B were assigned to x. Five of the syllables from List B were so assigned for any one *S*. Since there were 11 blocks of pairings in all in Experiment II (8 during acquisition and 3 during testing), four of these x's were repeated once and one was repeated twice. The digits from 1 to 4 were randomly assigned to i, j, k, and l.

Unlike Experiment I, the nature of response recording made it necessary to permit a failure to respond to test compounds.

RESULTS AND DISCUSSION

The summary of test responding given in Table 7 is like that in Table 2 except for the addition of a "no response" column. Each entry in this last column gives the total number of times *S*s failed to respond when the test figure of that row was shown. On the basis of the distribution of errors made to the training figures during the testing phase, a low-error group of 32 *S*s and a high-error group of 30 *S*s were defined, the range for the former being 0–7 and for the

TABLE 7
TEST RESPONSES OF ALL SUBJECTS IN EXPERIMENT II

Test figure	Response (and associated training cues)									
	i(*ab*)		*j*(*ac*)		*k*(*bc*)		*l*(*dx*)		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First
ad	11	7	8	3	6	2	96	48	3	2
bd	9	5	8	2	11	3	90	49	6	3
cd	6	3	6	3	18	9	89	46	5	1
Pooled	26	15	22	8	35	14	275	143	14	6

latter 8–17. The breakdown of test responses by error group is given in Table 8.

Since cues *a*, *b*, and *c* are symmetrical in the logical structure of the design (i.e., each having two associated responses during training), it is appropriate to pool the frequencies in each response category of Tables 7 and 8 over the three test figures. When this is done, we observe that of 372 responses (including failures) given by all *Ss* in the Both category, 275, or 74%, were instances of the response associated with cue *d* during training; the corresponding percentages for low- and high-error *Ss* were 82 and 66, respectively.

Since according to the component model only 50% of the test responses should have been instances of the response associated with cue *d*, it is clear that under the conditions of this experiment, *Ss*' preference for the relatively unambiguous cue was much more pronounced than in Experiment I. Owing to the design of the training series, this preference presumably could not be the result of *Ss* having learned to ignore the more ambiguous cues. However, in Experiment II, the unambiguous cue, *d*, differed from the other test cues also in that it had not been part of a regularly recurring pattern during training. Thus the possibility is suggested that *Ss* were, in effect, reinforced for responding to cue *d* whenever it appeared in a new compound. The next experiment is designed to rule out any such contingency.

EXPERIMENT III

The design of this experiment, summarized in Table 9 ensures that, as in Experiment II, *Ss* cannot be reinforced for selectively ignoring any of the test cues during training. Further, in the new paradigm, all

TABLE 8
TEST RESPONSES OF LOW- AND HIGH-ERROR SUBJECTS IN EXPERIMENT II

Test figure	Response (and associated training cues)									
	i(*ab*)		*j*(*ac*)		*k*(*bc*)		*l*(*dx*)		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First
Low-error *Ss*										
ad	3	3	2	1	3	2	55	25	1	1
bd	4	4	1	0	5	1	51	25	3	2
cd	1	1	5	2	6	3	51	25	1	1
Pooled	8	8	8	3	14	6	157	75	5	4
High-error *Ss*										
ad	8	4	6	2	3	0	41	23	2	1
bd	5	1	7	2	6	2	39	24	3	1
cd	5	2	1	1	12	6	38	21	4	0
Pooled	18	7	14	5	21	8	118	68	9	2

TABLE 9

DESIGN OF EXPERIMENT III

Training		Testing syllable combinations
Syllable combinations	Responses	
ab	*i*	*af*
ac	*j*	*ae*
ad	*k*	*ef*
eb	*l*	*bc*
ed	*m*	*ce*
fc	*n*	*cd*
		bf
		df

cues are components of regularly recurring patterns during training, and thus all should be equally likely to be sampled when they appear in test compounds. It was our intention that the only differentiation among the cues which might prove relevant to test behavior would occur along the dimension of ambiguity, or validity. To permit more thorough study of this variable, the relations between cues and reinforced responses were such that one cue (*a*) would be associated with three different responses, four cues (*b*, *c*, *d*, and *e*) with two different responses each, and one cue (*f*) with a single response. Then cues at each level of ambiguity were paired with cues at each of the other levels during the testing phase.

METHOD

Subjects. A total of 61 *S*s was used, all recruited from sections of the introductory course.

Apparatus. The arrangement of booths and the surfaces of the desks were as in Experiment II, but the projector-screen method of stimulus presentation and the switch-bank method of programming were not used. Instead, the stimuli were presented by means of digital display units manufactured by Industrial Electronics, Incorporated. Each such unit is capable of displaying any 1 of 12 different stimuli by having a bulb turned on behind a film containing the stimulus of interest; a lens arrangement bends the light so that all stimuli fall in the same location on the viewing surface. The sequence of stimulus events was programmed by a Friden tape reader and *S*s' responses were recorded by use of the Friden punch, as in Experiment II.

Procedure. The overall procedure discussed under Experiment I and Experiment II was followed with the schema shown in Table 9. There were 12 paired associates in each acquisition block and a total of eight acquisition blocks.

Another training block was given during the first part of the testing phase and then four more blocks with interspersed test compounds. There were 16 such tests and these were randomly interspersed under the restriction of exactly four tests per block. The syllables from List A (Experiment II) were randomly assigned to the letters *a* through *f* (Table 9) and the digits 1–6 to the letters *i* through *n*.

RESULTS AND DISCUSSION

Test responses for all *S*s are summarized in Table 10, and for low-error and high-error *S*s separately in Tables 11 and 12, respectively. The ranges of errors on training figures during the testing phase were 3–20 for the 30 low-error *S*s and 21–58 for the 31 high-error *S*s.

There is a total of 49 response failures among the 976 tests for all *S*s, giving a relative frequency of .050. The corresponding relative frequency for the low-error *S*s is .033. As can be seen in Tables 10 and 11,

TABLE 10

TEST RESPONSES OF ALL SUBJECTS IN EXPERIMENT III

Test figure	Response (and associated training cues)													
	i(*ab*)		*j*(*ac*)		*k*(*ad*)		*l*(*eb*)		*m*(*ed*)		*n*(*fc*)		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First
af	15	6	20	10	12	6	5	2	5	2	57	29	8	6
ae	27	12	16	9	12	8	27	14	27	14	5	2	8	2
ef	7	5	10	5	11	5	26	14	14	6	49	23	5	3
bc	24	10	18	8	17	9	24	11	7	5	23	12	9	6
ce	10	4	20	9	7	4	23	13	25	14	31	13	6	4
cd	9	5	15	10	29	16	7	5	24	9	30	12	8	4
bf	21	15	8	2	10	5	15	7	9	5	56	26	3	1
df	6	2	3	2	22	11	9	6	18	8	62	30	2	2

TABLE 11

TEST RESPONSES OF LOW-ERROR SUBJECTS IN EXPERIMENT III

Test figure	Response (and associated training cues)													
	i(*ab*)		*j*(*ac*)		*k*(*ad*)		*l*(*eb*)		*m*(*ed*)		*n*(*fc*)		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First
af	9	3	9	4	3	2	1	0	1	1	35	18	2	2
ae	8	4	9	4	4	3	19	9	18	9	0	0	2	1
ef	1	1	5	3	4	2	13	6	4	2	32	15	1	1
bc	16	6	10	5	8	4	8	4	2	2	13	7	3	2
ce	4	1	9	4	3	2	14	7	14	8	15	7	1	1
cd	3	2	5	3	20	13	2	1	9	4	17	5	4	2
bf	14	9	1	0	3	2	5	2	2	1	33	15	2	1
df	1	1	0	0	17	10	2	2	10	4	29	12	1	1

the distribution of response failures over test figures is rectangular, except perhaps for the compounds in which *f* is paired with a cue of intermediate ambiguity (*e*, *b*, *d*). But even in such cases the slight differences would make it reasonable to assume, when testing component model predictions, that the likelihood of response failure is adequately represented by a single probability common to all test figures for a given *S* grouping.

For convenience in discussing the various tests, we shall refer to *a* as the high-ambiguity (H) cue, *f* as the low-ambiguity (L) cue, and *b*, *c*, *d*, and *e* as intermediate ambiguity (I) cues. Test figures involving the same ambiguity combinations will be grouped together for analysis:

H-L : *af*

H-I : *ae*

I-L : *ef*, *bf*, *df*

I-I : *bc*, *ce*, *cd*

Further, we shall designate appropriate responses to a test figure (those reinforced to either component during training) by the letter A, other responses by O, and response failures by N. Within the A class, + and − will denote the responses associated with the less and more ambiguous cues, respectively.

We may quickly dispose of the I-I category to which little theoretical interest attaches. As expected, the division of A responses between the two cues was approximately equal, 48%, over all *S*s, being instances of *j* or *n*, the responses associated with cue *c*. For some reason not obvious to us, within this class, the *S*s showed a rather marked preference (84 *n*'s to 53 *j*'s) for the response associated with the less ambiguous

TABLE 12

TEST RESPONSES OF HIGH-ERROR SUBJECTS IN EXPERIMENT III

Test figure	Response (and associated training cues)													
	i(*ab*)		*j*(*ac*)		*k*(*ad*)		*l*(*eb*)		*m*(*ed*)		*n*(*fc*)		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First
af	6	3	11	6	9	4	4	2	4	1	22	11	6	4
ae	19	8	7	5	8	5	8	5	9	5	5	2	6	1
ef	6	4	5	2	7	3	13	8	10	4	17	8	4	2
bc	8	4	8	3	9	5	16	7	5	3	10	5	6	4
ce	6	3	11	5	4	2	9	6	11	6	16	6	5	3
cd	6	3	10	7	9	3	5	4	15	5	13	7	4	2
bf	7	6	7	2	7	3	10	5	7	4	23	11	1	0
df	5	1	3	2	5	1	7	4	8	4	33	18	1	1

of the two cues paired with c during training.

Considering now the A responses made by all *Ss* on tests involving differential ambiguities, for the H-L tests 55% of these responses were in the A+ category, for H-I, 50%, and for I-L, 59%. Corresponding values for the low-error *Ss* were 62%, 64%, and 60%, respectively. There would, thus, seem to be some preference for the less ambiguous cues, particularly among the low-error *Ss*. However, these percentages are not directly comparable with each other or with an expectation of 50% since an a priori expectation of 50%+ and 50%− responses within the A category is based on the assumption that all responses are determined by *Ss'* learning histories relative to the test cues. Since there were, in fact, incidences of inappropriate responses and response failures, adjusted theoretical values for the component model must be computed.

Except for the necessity of dealing with the N category, the analysis proceeds as in the case of Experiment I. For the data of all *Ss*, in the Both category, we have already estimated the probability of a response failure on any test, p_N, to be .050. Letting p denote probability of sampling the relevant cues on test trials, as in Experiment I, we can set up the observation equations:

For H-L

$$.852 = .950\ [p + (1 - p)2/3],$$

for H-I

$$.893 = .950\ [p + (1 - p)5/6],$$

for I-L

$$.773 = .950\ [p + (1 - p)1/2],$$

and for I-I

$$.781 = .950\ [p + (1 - p)2/3].$$

In each instance, the quantity on the left is the observed proportion of A responses and the factor .950 on the right is $1 - p_N$. The multiplier of $(1 - p)$ in each equation is the probability that a response occurring by chance (i.e., not determined by the test cues) will fall in the A category. Adding these equations (with proper weighting) and solving for p, we obtain the estimate $\hat{p} = .581$, and using this value we arrive at the component model predictions for the differential ambiguity tests.

$$\text{H-L} : P(\text{A}+) = .950\ [.581/2 + .419/6] = .342$$

$$\text{H-I} : P(\text{A}+) = .950\ [.581/2 + .419/3] = .408$$

$$\text{I-L} : P(\text{A}+) = .950\ [.581/2 + .419/6] = .342,$$

to be compared with corresponding observed values of .467, .442, and .456, respectively. Thus, the deviations from the component model occur in the expected order from largest to smallest: H-L, I-L, H-I.

A similar analysis for low-error *Ss* only yields a much larger estimate of the probability of sampling the relevant cues on test trials, $\hat{p} = .751$, and component model predictions of

$$\text{H-L} : P(\text{A}+) = .404$$

$$\text{H-I} : P(\text{A}+) = .445$$

$$\text{I-L} : P(\text{A}+) = .404,$$

to be compared with observed proportions of .583, .617, and .522, respectively. The order in this case is: H-L, H-I, I-L, with little difference between the first two.

It is difficult to support all of the comparisons of interest with formal significance tests, owing to the multiple observations per *S*. However, χ^2 (on the conditional space) values for the comparison of A+ and A− frequencies in the First data with component model predictions are significant at the .05 level or beyond for H-L and I-L comparisons. Considering also the uniformity of the deviations from the component model base line, one can scarcely avoid concluding that our *Ss* tended to base their test responses on the less ambiguous cue in the test figure. Comparison of these results with those of the Experiment II suggests that this preference is augmented when the training conditions provide differential rein-

forcement for responding to an unambiguous cue, but that the preference may be substantial even in the absence of such contingencies.

EXPERIMENT IV

It will be recalled that the test series of Experiment I included a compound *ad*, of which cue *a* had been associated with responses *i* and *j*, and cue *d* with responses *i* and *k* during training. We had anticipated, on intuitive rather than formal theoretical grounds, that *Ss* would exhibit a differential preference for the common response, *i*, on the *ad* test. Something of the sort appears to occur in many nonlaboratory situations. If, for example, a patient has several symptoms, *a*, *b*, *c*, ···, each of which is associated with two or more different diseases but all consistent with some one disease *X* (e.g., *a* is associated with *X* and *Y*, *b* with *X* and *W*, *C* with *X* and *Z* ···) one would surely expect a diagnostician to select *X* with high probability. Yet little evidence for any such effect was manifest in the data of Experiment I.

One possible interpretation is that selective response on the basis of a principle of response communality, or confirmation, is not a general characteristic of discrimination learning, but must be established in particular situations by special instructions or training. However, it is possible also that the conditions of Experiment I were unfavorable to manifestation of a response confirmation principle, perhaps because cues *a* and *d* were in a sense irrelevant to the discriminations learned during training. Thus we propose in the present experiment to arrange a more thorough test, using a training paradigm completely balanced with respect to cue validity and response frequency.

METHOD

Subjects and apparatus. Ninety-three *Ss* from the introductory psychology course were run in the same apparatus used in Experiment III.

Procedure. The training list included 16 paired-associate items, each presented once per block. The stimuli were syllable pairs conforming to the paradigm of Table 13 (with each pair appearing in both left-right orders). Stimulus lists A and B of Experiment II were pooled and eight syllables

TABLE 13
DESIGN OF EXPERIMENT IV

Training: Syllable combinations	Training: Responses	Testing syllable combinations
ab	*i*	*ad*
ac	*j*	*bc*
db	*k*	*ef*
dc	*i*	*gh*
eg	*l*	
eh	*j*	
fg	*k*	
fh	*l*	

selected for random assignment to the positions denoted by the letters *a–h* in Table 13. Responses were the digits 1, 2, 3, and 4, randomly assigned to the letters *i–l* in Table 13. The pairing of syllable combinations with responses was such that the list could not be completely learned if any of the cues were selectively ignored during training. Further, each component cue was associated with exactly two responses during training, and the test combinations were so chosen that each test provided an opportunity for operation of any tendency to respond on the basis of response communality.

The learning phase comprised 12 blocks, with the 16 items occurring in random order in each block. At the start of the testing phase, one additional training block was given; then the eight test compounds were interspersed through two training blocks, the positions being random except that exactly four tests occurred in each block.

RESULTS AND DISCUSSION

Full-test data are given for all *Ss* in Table 14 and for low- and high-error subgroups (with 47 and 46 *Ss*, respectively) in Table 15. Since all test compounds are symmetrical with respect to the experimental design, the data can conveniently be pooled, using theoretically significant response categories as in Experiment III. In this instance, the category A+ will denote the appropriate response to a test compound which is "correct" according to the notion of response communality (response *i* to *ad*, *l* to *ef*, etc.), A− any other appropriate response, O an inappropriate response, and N a response failure.

Taking first the pooled frequencies for all *Ss* shown in the upper portion of Table 16, we obtain directly $p_N = 35/744 = .047$ as our estimate of the probability of a response failure on any trial. Then, entering

TABLE 14

TEST RESPONSES OF ALL SUBJECTS IN EXPERIMENT IV

Test figure	Response (and associated training cues)									
	i(ab-dc)		*j(ac-eh)*		*k(db-fg)*		*l(eg-fh)*		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First
ad	95	47	38	15	37	23	9	4	7	4
bc	75	37	51	25	40	21	11	4	9	6
ef	12	9	30	17	49	25	87	38	8	4
gh	15	8	40	20	33	15	87	45	11	5

the appropriate values in

$$P(\mathrm{A}) = (1 - p_{\mathrm{N}})[p + (1 - p)3/4],$$

we obtain

$$\frac{662}{744} = .953\,[p + .75\,(1 - p)],$$

which yields the estimate $\hat{p} = .736$ for the probability that any test response is determined by the relevant cues. Using these estimates, we have computed the component model predictions given in the upper portion of Table 16. The same analysis for low-error *Ss* yields the parameter estimates $\hat{p}_{\mathrm{N}} = .056$, $\hat{p} = .844$, and the component model predictions presented in the lower portion of Table 16.

On the basis of the comparisons planned in advance of the experiment, i.e., those available in Table 16, we must conclude that there are only slight and statistically insignificant deviations from the component model base line in the direction of a preference for the A+ category. On an a posteriori basis, however, an interesting observation emerges regarding the trend over the test series. Since, as may be seen in Table 16, the response frequencies in the First category fell notably closer to component model predictions than did those in the Both category, the Second test responses must have deviated further. Examining the Second category for all *Ss* (obtainable by subtraction in Table 16) we find 177 A+ responses compared to 154 predicted; and, for the observed theoretical comparison over the A+ and A− entries only, we obtain a χ^2 of 5.82 which is significant at the .02 level. A similar analysis for low error *Ss* yields a χ^2 of 5.93. Thus, although *Ss*' behavior on early tests conforms closely to the component model, it is possible that a tendency to respond in terms of response communality develops as a function of continued experience with admixed training and test trials.

TABLE 15

TEST RESPONSES OF HIGH- AND LOW-ERROR SUBJECTS IN EXPERIMENT IV

Test figure	Response (and associated training cues)									
	i(ab-dc)		*j(ac-eh)*		*k(db-fg)*		*l(eg-fh)*		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First
Low-error *Ss*										
ad	51	28	19	8	20	10	1	1	3	0
bc	37	16	29	16	19	9	3	1	6	5
ef	3	3	16	10	22	13	47	18	6	3
gh	7	3	18	11	14	6	49	23	6	4
High-error *Ss*										
ad	44	19	19	7	17	13	8	3	4	4
bc	38	21	22	9	21	12	8	3	3	1
ef	9	6	14	7	27	12	40	20	2	1
gh	8	5	22	9	19	9	38	22	5	1

TABLE 16

POOLED TEST RESPONSES OF EXPERIMENT IV COMPARED WITH PREDICTIONS FROM COMPONENT MODEL

Category	Response type							
	A+		A−		O		N	
	Observed	Theoretical	Observed	Theoretical	Observed	Theoretical	Observed	Theoretical
All *S*s								
Both	344	308	318	354	47	47	35	35
First	167	154	161	177	25	23	19	18
Low-error *S*s								
Both	184	164	157	177	14	14	21	21
First	85	82	83	88	8	7	12	11

EXPERIMENT V

In virtually all studies of discrimination learning conducted with the classical (*ab* − 1; *ac* − 2) design, and in the first three experiments of the present series, the factor of cue ambiguity, or validity, has been unobtrusively but ubiquitously confounded with stimulus frequency. Consider, for example, the training paradigm of Experiment III (Table 9). Cue *f*, which proved to be the most potent controller of responding on transfer tests, occurred only once per training block, whereas the most ambiguous cue, *a*, occurred three times, and those of intermediate ambiguity twice per block.

As a step toward disentangling this confounding, we propose in this experiment to maintain the same arrangements of training cues and responses as in Experiment III, and to use the same transfer tests, but to change the relative frequencies of occurrence of some of the stimulus-response combinations during training. Thus we shall be able to compare test responding to cues which differ in ambiguity with training frequencies equated; and to compare cues which are of similar ambiguity but differ with respect to training frequency.

METHOD

Subjects and apparatus. The 144 *S*s all were students in introductory psychology. Apparatus was the same as that of Experiment III.

Procedure. Training conditions are summarized in Table 17. The new feature is that, whereas in Experiment III each syllable combination and its associated response occurred once per training block, in the present experiment some combinations occurred more often than others. For Conditions F− (69 *S*s), and F+ (75 *S*s) there were 20 and 26 trials, respectively, per training block, as compared to 12 in Experiment III, each left-right arrangement of syllable combinations being assigned the frequency per block shown in Table 17. In Condition F−, any effects of training frequency should be reduced for tests involving the least ambiguous cue, *f*, whereas in Condition F+ they should be accentuated. The F− condition will provide instances both of tests with differential ambiguity but equal training frequency (e.g., *ef*) and of tests with similar ambiguity but unequal training frequency (e.g., *bc*). The F+ condition will provide tests of the second type (e.g., *ce*) and also tests on which effects of training frequency and ambiguity are strongly confounded (e.g., *af*).

Training and testing were conducted as in Experiment III, except for the variations in makeup of training blocks. In the first training block each syllable pair appeared once, the two left-right arrangements occurring back-to-back. Then the differential stimulus frequencies shown in Table 17 applied throughout the subsequent eight training blocks, and throughout the four additional blocks of the testing phase, in each of which four test trials were interspersed among the training trials. In all blocks after the first, the order of stimulus presentation was entirely random. The

TABLE 17

TRAINING CONDITIONS OF EXPERIMENT V

Syllable combinations	Responses	Frequency of occurrence per block	
		Condition F−	Condition F+
ab	*i*	4	4
ac	*j*	1	4
ad	*k*	1	2
eb	*l*	1	1
ed	*m*	1	1
fc	*n*	2	1

syllables of List A (Experiment II) were assigned to the letters *a* through *f* in Table 17 and the digits 1–6 to the letters *i* through *n*.

Results and Discussion

Frequencies of test responses following the F− training condition are summarized in Table 18 and those following the F+ condition in Table 19. Similarly, Tables 20 and 21 show the frequencies of test responses for the low-error *S*s (34 in F− and 37 in F+) while Tables 22 and 23 show these frequencies for high-error *S*s (35 in F− and 38 in F+). As usual, deviations from the component model base line, and, in particular, differences between A+ and A− frequencies, were greater for low-error than for high-error *S*s. However, all of the principal trends to be discussed hold for both subgroups, so only the combined data for all *S*s will be used in subsequent discussion and comparisons.

Direct comparisions of F− and F+ conditions with each other and with Experiment III (which might be denoted as an F= condition) can most conveniently begin by reference to Table 24. Here, test response frequencies over all *S*s in Experiment III and in both conditions of Experiment V are shown jointly in percentage form (for the Both category). To facilitate theoretical comparison, the data have been combined further in Table 25, where test response percentages for each of the response types defined in the analysis of Experiment III are given. The pooling in Table 25 is over all figures representing each differential ambiguity combination.

TABLE 18

Test Responses of All Subjects in Experiment V, Condition F−

Test figure	Response (and associated training cues)													
	i(ab)		*j(ac)*		*k(ad)*		*l(eb)*		*m(ed)*		*n(fc)*		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First
af	11	9	27	14	23	8	9	4	4	2	49	28	15	4
ae	16	8	13	4	16	9	52	25	23	12	4	4	14	7
ef	2	1	7	4	13	9	37	19	26	12	44	19	9	5
bc	21	11	35	17	11	3	18	10	7	4	33	15	13	9
ce	6	3	25	12	10	5	36	16	24	14	29	16	8	3
cd	2	1	26	14	31	15	16	9	30	13	26	13	7	4
bf	26	16	17	5	11	4	17	11	4	3	54	25	9	5
df	3	2	4	2	32	13	9	4	27	17	48	22	15	9

TABLE 19

Test Responses of All Subjects in Experiment V, Condition F+

Test figure	Response (and associated training cues)													
	i(ab)		*j(ac)*		*k(ad)*		*l(eb)*		*m(ed)*		*n(fc)*		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First
af	10	6	12	6	10	6	3	3	2	1	108	51	5	2
ae	14	4	12	6	19	7	56	29	39	24	2	0	8	5
ef	3	1	3	1	2	1	19	12	20	10	101	49	2	1
bc	42	17	49	23	7	2	22	13	7	5	21	13	2	2
ce	4	3	41	24	6	3	45	26	27	7	19	8	2	4
cd	5	0	31	12	57	33	7	4	27	15	13	5	8	4
bf	26	14	4	3	2	0	9	4	5	2	100	49	10	6
df	6	4	2	1	18	8	2	0	12	7	101	48	4	3

TABLE 20

TEST RESPONSES OF LOW-ERROR SUBJECTS IN EXPERIMENT V, CONDITION F−

Test figure	Response (and associated training cues)													
	i(ab)		*j(ac)*		*k(ad)*		*l(eb)*		*m(ed)*		*n(fc)*		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First
af	4	3	12	7	9	3	2	1	1	0	33	19	7	1
ae	5	3	5	2	6	3	30	13	12	6	3	3	7	4
ef	0	0	3	3	1	1	20	9	10	6	31	13	3	2
bc	10	6	20	7	4	2	10	7	1	1	19	9	4	2
ce	2	1	17	7	2	1	21	10	8	5	15	9	3	1
cd	0	0	17	10	13	5	5	4	16	7	14	6	3	2
bf	10	7	9	2	1	0	9	6	2	2	33	15	4	2
df	0	0	1	1	13	5	3	2	13	9	31	14	7	3

TABLE 21

TEST RESPONSES OF LOW-ERROR SUBJECTS IN EXPERIMENT V, CONDITION F+

Test figure	Response (and associated training cues)													
	i(ab)		*j(ac)*		*k(ad)*		*l(eb)*		*m(ed)*		*n(fc)*		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First
af	1	1	5	2	6	5	1	1	0	0	60	28	1	0
ae	7	2	6	4	9	2	32	17	18	10	0	0	2	2
ef	1	0	0	0	0	0	11	6	4	3	57	27	1	1
bc	23	9	22	11	4	2	11	6	1	1	13	8	0	0
ce	0	0	18	10	4	3	27	14	12	3	10	5	3	2
cd	2	0	9	6	31	17	2	1	20	10	7	1	3	2
bf	10	4	0	0	0	0	4	3	1	0	59	30	0	0
df	2	1	1	1	7	2	1	0	2	2	56	27	5	4

TABLE 22

TEST RESPONSES OF HIGH-ERROR SUBJECTS IN EXPERIMENT V, CONDITION F−

Test figure	Response (and associated training cues)													
	i(ab)		*j(ac)*		*k(ad)*		*l(eb)*		*m(ed)*		*n(fc)*		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First
af	7	6	15	7	14	5	7	3	3	2	16	9	8	3
ae	11	5	8	2	10	6	22	12	11	6	1	1	7	3
ef	2	1	4	1	12	8	17	10	16	6	13	6	6	3
bc	11	5	15	10	7	1	8	3	6	3	14	6	9	7
ce	4	2	8	5	8	4	15	6	16	9	14	7	5	2
cd	2	1	9	4	18	10	11	5	14	6	12	7	4	2
bf	16	9	8	3	10	4	8	5	2	1	21	10	5	3
df	3	2	3	1	19	8	6	2	14	8	17	8	8	6

TABLE 23

TEST RESPONSES OF HIGH-ERROR SUBJECTS IN EXPERIMENT V, CONDITION F+

Test figure	Response (and associated training cues)													
	i(ab)		*j(ac)*		*k(ad)*		*l(eb)*		*m(ed)*		*n(fc)*		No response	
	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First	Both	First
af	9	5	7	4	4	1	2	2	2	1	48	23	4	2
ae	7	2	6	2	10	5	24	12	21	14	2	0	6	3
ef	2	1	3	1	2	1	8	6	16	7	44	22	1	0
bc	19	8	27	12	3	0	11	7	6	4	8	5	2	2
ce	4	3	23	14	2	0	18	12	15	4	9	3	5	2
cd	3	0	22	6	26	16	5	3	7	5	6	4	7	4
bf	16	10	4	3	2	0	5	1	4	2	41	19	4	3
df	4	3	1	0	11	6	1	0	10	5	45	21	4	3

TABLE 24

SUMMARY OF TEST RESPONSES IN EXPERIMENTS III AND V IN PERCENTAGE FORM (ALL SUBJECTS, BOTH RESPONSES)

Test figure	Response (and associated learning stimuli)																				
	i(ab) Condition			*j(ac)* Condition			*k(ad)* Condition			*l(eb)* Condition			*m(ed)* Condition			*n(fc)* Condition			No response Condition		
	F=	F−	F+	F=	F−	F+	F=	F−	F+	F=	F−	F+	F=	F−	F+	F=	F−	F+	F=	F−	F+
af	12.3	8.0	6.7	16.4	19.6	8.0	9.8	16.7	6.7	4.1	6.5	2.0	4.1	2.9	1.3	46.7	35.5	72.0	6.6	10.9	3.3
ae	22.1	11.6	9.3	13.1	9.4	8.0	9.8	11.6	12.7	22.1	37.7	37.3	22.1	16.7	26.0	4.1	2.9	1.3	6.6	10.1	5.3
ef	5.7	1.4	2.0	8.2	5.1	2.0	9.0	9.4	1.3	21.3	26.8	12.7	11.5	18.8	13.3	40.2	31.9	67.3	4.1	6.5	1.3
bc	19.7	15.2	28.0	14.8	25.4	32.7	13.9	8.0	4.7	19.7	13.0	14.7	5.7	5.1	4.7	18.8	23.9	14.0	7.4	9.4	1.3
ce	8.2	4.3	2.7	16.4	18.1	27.3	5.7	7.2	4.0	18.9	26.1	30.0	20.5	17.4	18.0	25.4	21.0	12.7	4.9	5.8	5.3
cd	7.4	1.4	3.3	12.3	18.8	20.7	23.8	22.5	38.0	5.7	11.6	4.7	19.7	21.7	18.0	24.6	18.8	8.7	6.6	5.1	6.7
bf	17.2	18.8	17.3	6.6	12.3	2.7	8.2	8.0	1.3	12.3	12.3	6.0	7.4	2.9	3.3	45.9	39.1	66.7	2.5	6.5	2.7
df	4.9	2.2	4.0	2.5	2.9	1.3	18.0	23.2	12.0	7.4	6.5	1.3	14.8	19.6	8.0	50.8	34.8	67.3	1.6	10.9	6.0

TABLE 25

TEST RESPONSE PERCENTAGES IN EXPERIMENTS III AND V POOLED OVER FIGURES OF EQUAL AMBIGUITY (ALL SUBJECTS, BOTH RESPONSES)

Ambiguity pairing	Condition	Response percentage			
		A+	A−	O	N
H-L	F=	46.7	38.5	8.2	6.6
	F−	35.5	44.2	9.4	10.9
	F+	72.0	21.3	3.3	3.3
H-I	F=	44.3	45.0	4.1	6.6
	F−	54.4	32.6	2.9	10.1
	F+	63.3	30.0	1.3	5.3
I-L	F=	45.6	31.7	20.0	2.7
	F−	35.3	39.9	16.9	8.0
	F+	67.1	23.1	6.4	3.3

TABLE 26

COMPARISON OF EFFECTS OF AMBIGUITY AND FREQUENCY OF TEST CUES

Constant ambiguity—Unequal training frequency

Test figure	Condition	Frequency ratio	Response percentage			
			Less frequent A	More frequent A	O	N
cd and *ce*	F−	3:2	43.8	38.4	12.3	5.4
bc	F−	5:3	49.3	28.3	13.0	9.4
cd	F+	5:3	56.0	29.3	8.0	6.7
ce	F+	5:2	48.0	40.0	6.7	5.3
All above figures combined			48.3	34.9	10.4	6.4

Differential ambiguity—Equal training frequency

Test figure	Condition	Ambiguity grouping	Response percentage			
			A+	A−	O	N
ef and *df*	F−	I-L	33.3	44.2	13.8	8.7

Table 25 provides several comparisons in which ambiguity is constant while training frequency varies, and in every case the frequencies of test responses in the A categories exhibit a preference for the cue in the test figure which had the *lower* relative frequency during training. Thus, in the H-L tests, the A+ preference manifest for Experiment III is not evident at all in the F− condition of Experiment V, and the preference is greatly accentuated when the relative frequency of the H to the L cue (*a* to *f*) goes from 3:1 to 10:1. In the H-I tests, preference for A+ increases uniformly as the training ratio of H to I (*a* to *e*) goes from 3:2 in Experiment III to 6:2 in Experiment V, Condition F− and 10:2 in Experiment V, Condition F+. In the I-L tests, the moderate A+ preference shown in Experiment III disappears when the training ratios are changed toward equality in the F− condition of Experiment V, but is greatly increased when the ratios of I to L increase in the F+ condition.

Table 26 contains the remaining comparisons discussed earlier in the exposition of Experiment V. Included are the cases in which the test cues are of the same ambiguity but unequal in training frequency and in which the test cues are of differential ambiguity but equal in training frequency. With respect to effects of variation in training frequency when ambiguities are constant, these analyses are in accord with the relationship apparent in Table 25, although in some instances (e.g., the F+ data in Table 26) the degree of preference for the less frequent cue is not directly related to the training ratio. The one comparison available with ambiguity varied while training frequencies were held constant, the F− data for the test figures *ef* and *df* (Table 26), yields a higher percentage of A− than A+ responses. While this difference may not be significant it certainly indicates no trace of a preference for the low ambiguity cue.

EXPERIMENT VI

The results of Experiment V serve to cast considerable doubt on the hypothesis that transfer to new stimulus compounds is mediated by a principle of cue selection on the basis of minimizing ambiguity. Substantial deviations from component model predictions occurred when ambiguity was confounded with training frequency, but disappeared when the confounding was eliminated. However, the ordering of test response proportions in relation to training frequency was not entirely consistent, and only one comparison was available with cues having equal training frequencies but different degrees of ambiguity. The present experiment, utilizing the simplified design summarized in Table 27, will permit further

TABLE 27

TRAINING AND TESTING CONDITIONS OF EXPERIMENT VI

Training				Testing syllable combinations
Syllable combinations	Responses	Relative frequencies F=	F−	
ab	*i*	1	1	*ad*
ac	*j*	1	1	*cb*
dc	*k*	1	2	*bd*

evaluation of the effects of each of these variables singly, that is, with the other held constant. By reducing the number of training syllable combinations, we contrived also to reduce the incidence of test response failures to negligible proportions.

METHOD

Subjects. The *S* pool was again the students in introductory psychology; 104 were used in the F= condition and 99 in the F− condition.

Apparatus. Between Experiments V and VI, the laboratory was completely rebuilt concomitant with a shift in quarters. Six new booths were built with acoustic tiles on the sides and ceiling. Each booth contained a plywood panel which *S* faced as he sat in the booth. At the top of this panel was a 4-inch two-way loudspeaker for communication between *S* and the experimenter (who was in an adjoining room). At the center of the panel was a plate upon which the display unit appropriate for a given experiment could be mounted. Just below this plate, resting on the surface of a desk, was a response-unit containing eight Plexiglas windows with a green light below each window and a push-button below each green light. The overall arrangement and operation of these response units were as described under Experiment II for the original units.

The booths were in a single line so that the six *Ss* run at a given sitting entered from the same corridor. Behind the array of booths on the opposite wall of the corridor, between Booths 3 and 4, was a loudspeaker through which the instructions came. The instructions to *Ss* were all contained on a two-channel magnetic tape which was run through a tape-deck and amplifier connected to the loudspeaker.

As in Experiment III, stimuli were presented by digital display units, programmed by a Friden tape reader, and responses were recorded on the tape of a Friden punch. However, the operation was automated in that stimuli were selected and randomized by a generalized IBM 709 program (written in Fortran), the information in the output cards from the computer was put on paper tape by a Friden Flexowriter, and the tape for a squad of *Ss* run at a given time completely determined the sequence of experimental events. Thus, in proper sequence the paper tape operative in the reader turned the tape recorder on for instructions, allowed a pause for questions, started the learning phase, sounded a warning buzzer, presented stimuli and responses, introduced a rest pause, started the testing phase, presented test stimuli when appropriate, introduced modifying instructions where necessary and even triggered the final statement that the experiment was completed. Some safety devices were introduced to insure proper experimenter identification of all groups, counter resetting, and stimulus counterbalancing.

Various auxiliary paper tape codes—stop, pause, tape recorder on, etc.—were entered as symbols in the basic symbol list. The various intervals were timed by ATC timers; at the end of a given timed period the input tape was searched again for further instructions. The end of a set of instructions from the tape recorder was accompanied by a signal which initiated further reading of the input paper tape.

Procedure. The general procedure of Experiments III, IV, and V was followed. However, each training block contained only three different syllable pairs and there were only three different test figures. Eight acquisition blocks preceded the testing phase. The test combinations were randomly interspersed among additional training trials during the final two of the three blocks of the testing phase, three occurring in each of these blocks.

The letters *a*, *b*, *c*, and *d* of the experimental paradigm (Table 27), were randomly replaced (for each squad of six *Ss*) by the nonsense syllables VOP, GAK, FUH, and TEF and the letters *i*, *j*, and *k* were randomly replaced by the digits 1, 2, and 3. Once a set of these randomizations was made, the set was used for one squad under the F= condition and one under the F− condition.

Each training stimulus (i.e., each syllable pair shown in Table 27 in each left-right arrangement) occurred once per block under Condition F= whereas differential frequencies of presentation were introduced in Condition F−. There were thus six trials per training block under F= and eight per block under F−. A set of randomized orders of presenting the six training pairs of Condition F= was used for each squad of *Ss*, and the same order was used for the corresponding squad of Condition F− (correspondence being defined by the same stimulus and response assignment) with the random insertion of the additional stimulus-response occurrences within each block.

RESULTS AND DISCUSSION

The basic test frequency data for all *Ss* are given in Table 28 and the percentages of occurrence of the principal response types arranged by combinations of cue ambiguity and training frequency, in Table

TABLE 28

TEST RESPONSES OF ALL SUBJECTS IN EXPERIMENT VI

Test figure	Response (and associated training cues)							
	i(ab)		*j(ac)*		*k(dc)*		No response	
	Both	First	Both	First	Both	First	Both	First
Condition F=								
ad	23	11	53	32	130	59	2	2
bd	89	44	20	12	94	44	5	4
cb	125	62	57	30	20	12	6	0
Condition F−								
ad	25	17	81	42	91	39	1	1
bd	106	48	29	15	60	33	3	3
cb	138	68	33	22	27	9	0	0

29. As before, A+ denotes the response appropriate to the low ambiguity cue of an I-L (intermediate-low) test combination and A− the responses appropriate to the other cue; in this experiment there are, of course, no entries in the O (other appropriate responses) category for I-L tests. In the case of the L-L tests, A+ represents the response appropriate to cue *b*, which has relative frequency of one per training block under both conditions, and A− the response appropriate to the other cue of the pair. For convenience, the ratio of training frequencies is given for each test pair, the test cue combinations under I-L being listed separately since their training ratios differ in the F− condition.

A priori predictions of test response percentages from the component model are simply 50% A+ and 50% A−, and since the incidence of response failures is very low, adjusted predictions would be essentially the same for I-L tests under both conditions. Some O responses occur on the L-L tests, indicating that despite the simplified conditions *S*s are responding to irrelevant aspects of the test situation in a significant proportion of cases. However, responses determined by extraneous factors would be expected to fall in the A+ and A− categories equally often.

The principal conclusion to be drawn from the pattern of test response proportions exhibited in Table 29 is that degree of preference for the A+ category is uniformly related to training ratio but, when this variable is controlled, test responding is not in the direction of minimum cue ambiguity, or validity. For I-L tests following F= training, there is an apparent preference for the low ambiguity cue (i.e., for the A+ response category), but when the relative training frequencies are equated (test figure *ad* following F− training) the preference disappears. Data from L-L tests conform closely to component model predictions when training frequencies are equal (F=) but exhibit a substantial preference for the less frequent cue (i.e., for the A+ response category) when they are unequal.

It is perhaps of some interest to note that the slightly higher percentage for A− rather

TABLE 29

TEST RESPONSE PERCENTAGES OF ALL SUBJECTS IN EXPERIMENT VI

Test type	Frequency ratio A+:A−	Response type[a]			
		A+	A−	O	N
Condition F =					
I-L					
ad	1:2	62.5	36.5	—	1.0
cb	1:2	60.1	37.0	—	2.9
L-L	1:1	42.7	45.2	9.6	2.4
Condition F −					
I-L					
ad	2:2	46.0	53.5	—	0.5
cb	1:3	69.7	30.3	—	0.0
L-L	1:2	53.5	30.3	14.6	1.5

[a] For the L-L tests, A+ denotes the response associated with Cue *b* and A− the response associated with *d*; for the I-L tests, A+ and A− denote responses associated with the low and intermediate ambiguity cues, respectively.

than A+ in test compound *ad*, Condition F−, is in accord with a similar finding for test figures *ef* and *df* of Experiment V, Condition F− (see Table 26). The difference between A+ and A− in both cases lies in the way the equal frequency per block was accomplished: for A+ the same learning compound was shown repetitively as a unit while A− was embedded in different learning compounds.

DISCUSSION

To recapitulate our principal findings: (*a*) Transfer to new compounds following training under a standard discrimination learning paradigm (involving two training patterns with overlapping components) was fairly well described by the component model. Some deviations from the test response probabilities predicted by the model, though not significant, suggested the possibility that *S*s tended to sample preferentially cues with relatively high validities or low ambiguities—that is, cues which had been the relatively best predictors of reinforcing events during training. (*b*) In a newly designed experiment with conditions arranged to facilitate any tendencies to respond in terms of cue ambiguities, *S*s exhibited very marked preferences for low ambiguity cues in test compounds. (*c*) Controlled comparisons provided by further experiments showed that, in many cases at least, the apparent preference for low ambiguity cues is actually a matter of preferring the test cue of highest "novelty."

One's first reaction to these findings is perhaps to ask why the deviations from component model predictions have not been conspicuous in previous studies of stimulus compounding. The principal answer is quite likely that, as a general rule, relative frequencies of the various cues which appear in the test compounds have been equal during training. This was the case, for example, in the well-known study by Schoeffler (1954), which provided the first clean-cut experimental demonstration of successful prediction of transfer results via the component model, and in the study of transfer effects following discrimination learning by Estes, Burke, Atkinson, and Frankmann (1957).

Next, one may well wonder why some of the auxiliary principles of transfer which have arisen independently in other contexts do not seem to enter in any important way into the determination of test behavior in our studies. The most intuitively compelling of these auxiliary principles is, perhaps, that of cue validity, which was indeed the foundation of Restle's theory of discrimination learning (1955). According to this concept, for which Restle (1955), in turn, acknowledges indebtedness to Lawrence (1950), *S* during the course of discrimination learning comes to ignore cues which are ambiguous, that is which are unreliable predictors of reinforcing events, and to sample preferentially those cues which are less ambiguous, that is, which are reliable predictors of reinforcing events. An aspect of the concept which was not formally stated by Restle, but which has been assumed in many applications, especially in the area of concept learning (Bourne & Restle, 1959) is that following such training *S*s should preferentially sample the cues of low ambiguity, or high validity, when they are encountered in transfer situations.

Taking, on the one hand the results of our Experiment II, in which very strong tendencies for responding in terms of the least ambiguous cue were manifest on transfer tests and, on the other, the results of the remaining experiments which show little or no effect of cue ambiguity when other variables are adequately controlled, we are inclined to conclude that a generalized tendency to sample selectively cues of low ambiguity does not spontaneously arise in the course of discrimination learning. However, it seems clear also that a tendency toward repeated selection of a particular cue may develop when such consistency is specifically reinforced during training. Evidently, conditions were near optimal in our Experiment II, in which the low-ambiguity cue was not part of a recurrent pattern during training but was continually re-paired with other, "transient," cues. It might be noted that the type of reinforcement contingency existing in this experiment is characteristic of experiments on concept identification, in the context of which the

hypothesis of selective sampling on the basis of cue validity has proved fruitful.

The notion of transfer responding in terms of response communality has been most clearly stated by Trabasso and Bower (1964) in connection with the interpretation of multiple category concept learning experiments. For example, in discussing the learning of a four-category problem, in which *S*s learn to assign correctly to one of four categories stimulus patterns which are either circular or triangular in form and either orange or blue in color, Trabasso and Bower (1964) state the principle of response communality as follows:

> We now wish to know the probability with which the *S* will give any one of the four responses to a particular pattern shown on trial *n* (e.g., an orange circle). The performance rule is this: the subject generates a pair of covert responses for each relevant attribute and then overtly responds with the common element (intersection) from these two sets [p. 145].

Again, on the basis of our Experiment IV, we are inclined to conclude that this type of performance tendency does not spontaneously develop in the course of ordinary discrimination training, though it may under the reinforcement contingencies of concept identification studies.

The one auxiliary principle which we have found to operate ubiquitously, and evidently to involve no special conditions of differential reinforcement for its appearance, is that of inverse frequency, or to put it more positively, *relative novelty*. Almost unfailingly throughout our entire series of experiments, *S*s have been found to sample preferentially from test compounds the cues which had occurred less frequently during previous training. We failed to appreciate the power of this effect until relatively late in our series of studies because not only in the standard discrimination paradigm, but also in many of the modified designs used in our experiments, there is a negative correlation between cue validity and relative frequency. That is, cues of higher validity tend to occur less frequently than cues of lower validity during training simply as a consequence of the fact that the less valid, or more ambiguous, cues tend to be common to two or more training patterns whereas a cue of low ambiguity, or high validity, ordinarily belongs to only a single training pattern. However, although we were not, in fact, prepared for the importance of the relative novelty principle prior to our experiments, it appears in hindsight that we might have been, for this principle has forced its way to the attention of investigators in other experimental contexts. In fact the principle of relative novelty is an important component of Broadbent's (1958) filter theory:

> ...responses to a novel stimuli are in fact particularly efficient: it is worth digressing at this stage to consider the evidence on this topic.
>
> A particularly good example was given in Chapter 2, when describing the experiments of Poulton (1956) on listening to messages from several loudspeakers. If the messages came more frequently from one loudspeaker, a message from the previously quiet speaker was more likely to be correctly heard than a message from the previously busy speaker. In terms of the filter theory we put forward earlier, the filter is biassed towards previously quiet channels, and information on busy channels has a lower chance of reaching the perceptual system. In ordinary speech, we attend to an unusual event rather than a simultaneous usual event. This fact is particularly curious because a rare event contributes more information than a common event, on the measure considered in Chapter 3: and a task requiring the analysis of more information might be expected to be more difficult. It seems well supported by experiments in other fields, however. For example, Hyman (1953), working with visual reaction times, found that as the ensemble of signals increased the reaction time went up, confirming the finding by Hick (1952) that reaction time was proportional to the information in the signal. When Hyman altered the frequency of the signals instead of having them equiprobable, he found once again that this decrease in the average information per signal did give a decrease in the average reaction time. But reactions to the least common signals were faster than they should have been on information calculations. Once again there seemed to be an undue bias in favour of the unusual event.
>
> Berlyne (1951a) also used visual reactions but required the subject to react only to one out of a group of simultaneous signals; and the most frequently chosen signal was recorded. After a sequence of similar groups, if a group was presented in which all members but one were familiar, the unusual signal was the most likely to be chosen. The same author, using rats (1950), has introduced the animals to particular objects, then removed them, and faced them later with some of the former objects and a new one. The animals spent more time investigating the new object than the previously experienced ones. As he points out, the

time scale is in this case different, but once again the unusual stimulus is more likely to elicit a response.

If the response to a fresh stimulus is particularly efficient, it does not seem plausible that the inefficiency of work, done just after a noise has been turned on or off, should be due to a general decline in ability to respond, due to some competing startle response. Such a competing response should interfere with responses to the novel stimulus itself as well as with responses to other stimuli. It seems rather more likely that the man is unable to respond to visual stimuli from his task because he is taking in information from the ear; so he would actually be more efficient on responses to the auditory stimulation. Our view of the events within the man would be as follows.

His capacity is limited, and therefore a filter placed early in his nervous system selects only part of the information reaching his sense-organs. This will normally represent information necessary for his task. The filter has a bias, however, towards channels on which any novel event occurs [pp. 84–86].

The influence of novelty may even be evident in the preference for A− over A+ when both are shown equally frequently per learning block (see Table 26, test figures *ef* and *df*, and Table 29, test figure *ad* under Condition F−). The equality of frequency was achieved by repetition of the full learning compound containing A+, which implies a slight edge toward A− in novelty.

Perhaps some emphasis should be placed upon the distinction between the novelty principle and the matching rule. Matching was illustrated in the introduction to this report by findings of Binder and Feldman (1960). Thus, for example, if the pairing *ac*-A_1 occurred twice as often per block as the pairing *ad*-A_2, matching leads to the expectation that on a test with the single cue *a*, responses A_1 and A_2 should occur in a ratio of close to 2 to 1. While the novelty principle refers to cue selection when the stimulus is a compound, matching refers to response selection when alternative responses are appropriate for a particular cue.

The process of cue selection on the basis of novely has not been an issue in experiments aimed at establishing the conditions for matching since compounds have not typically been involved in the test configurations. One exception lies in the research of Feldman (1963) which included within a larger array, test compounds requiring cue selection. However, even Feldman's research is of little relevance for our principle of novelty since the phase preceding his test trials involved concept rather than discrimination learning.

It appears that for the general class of transfer situations under consideration in this study the component model requires augmentation by a novelty principle, but we do not have immediately at hand any rational basis for a general quantitative formulation. A simple first approximation would be to assume simply that when cues from the training patterns of a discrimination learning series are recombined in a transfer test, the sampling probability of each test cue is equal to the probability that it was not the last to appear on training trials preceding the test. Thus, if the relative frequencies of cues *a* and *b* during training were in the ratio x:y, the probability that cue *a* would be sampled from the test compound *ab* would be $y/(x + y)$.

To gain an idea as to how well this provisional form of an augmented component model accounts for effects in our data that appear to reflect the relative novelty principle, let us return to the data for the low-error *S*s of Experiment I (the first instance, in our study, of a significant deviation from component model predictions). Cues *a* and *e* occurred with frequencies in a 2:1 ratio during training. The estimate of .861 obtained previously for *p*, the probability that *S* samples *any* of the cues presented on a transfer test, is unaffected by the new assumption. Thus the probability of response *i* to test figure *ae*, is given by

$$\begin{aligned} &p\left(\frac{2}{3} + \frac{1}{3}\cdot\frac{1}{2}\right) + (1 - p)1/3 \\ &= .86\ (5/6) + .14/3 \\ &= .764. \end{aligned}$$

since, when *S* responds to the test cues he samples *e* with probability 2/3, and in that event certainly makes response *i*; and he samples *a* with probability 1/3, and in that event makes response *i* with probability 1/2. Proceeding similarly for the other responses, and converting predicted response

probabilities to frequencies, we obtain 62, 16, and 4 as the predicted frequencies of responses *i*, *j*, and *k* to be compared to the observed values 67, 12, and 3, respectively. For test figure *dc*, the predicted frequencies prove to be 16, 50, and 16 for *i*, *j*, and *k*, corresponding to observed values of 19, 53, and 10, respectively. The remaining theoretical entries in Table 5 are unaffected by the addition of the inverse frequency rule.

In Experiment II, the training frequencies of cue *d* and the cue paired with it on a transfer test were in each case in the ratio 2:1, whereas the frequencies, for low-error *S*s, of test responses appropriate to cue *d* and those appropriate to the cue paired with it were in a ratio of approximately 6.3:1 (Table 8). Clearly, as we had concluded on the basis of qualitative considerations earlier, the reinforcement contingencies of this experiment generate a sampling preference for the unambiguous cue far too extreme to be accounted for by the inverse frequency rule.

For the data of Experiment III, the augmented component model again proves quite adequate. Considering the data for the low-error *S*s, in which substantial excesses of A+ responses occurred for all types of transfer tests (see discussion above), and using the same *p* estimate as before, we obtain predicted A+ proportions of .585, .516, and .524 for H-L, H-I, and I-L tests, respectively, corresponding to observed values of .583, .617, and .522.

The best quantitative test of the augmented model is provided by Experiment VI. The number of different training patterns was small enough so that the simple form of the inverse frequency rule might be expected to hold to a good approximation, but at the same time, owing to the training conditions, a considerable variety of training-test relationships is available.

In this instance, even the a priori predictions, utilizing no parameters estimated from the data, come off rather impressively. For the F= condition, the predicted proportion of A+ responses is .677 for both test figures *ad* and *cb*, and .500 for *bd* (in which case A+ is the response associated with *b*); the observed proportions (excluding the O and N categories) are .631, .619, and .486 for the three figures respectively. For the F− condition, a priori predictions of .500, .750, and .667 for test figures *ad*, *cb*, and *bd* may be compared to the observed values .462, .697, and .639.

These results may serve to indicate that the inverse frequency principle, even in a very provisional quantitative form, corrects the main disparities between predictions from the component model and our data for various test conditions. Although somewhat better fits to our present data than those shown above could be obtained by systematically reestimating the various parameters, it does not seem worthwhile to push on in this direction. For one thing, the quantitative properties of the relative novelty principle need further specification. It seems almost certain that the weighting factor associated with the cue which represents its relative novelty must change as some systematic function of time since last occurrence. By analogy with related processes that have been dealt with in stimulus sampling theory, it seems likely that an exponential decay function will be called for (see Atkinson & Estes, 1963, pp. 219–223). It will probably prove expedient to undertake development of the more elaborate model in connection with suitably designed new experiments in which time intervals between occurrences of a given cue on training and testing trials can be controlled according to simpler schemes than those obtaining in the present study. From our present findings we conclude only that a relative novelty principle has considerable support at a qualitative level and is evidently a variable requiring careful attention in the design and interpretation of all types of transfer studies.

REFERENCES

ARCHER, E. J. A re-evaluation of the meaningfulness of all possible CVC trigrams. *Psychological Monographs*, 1960, **74**(10, Whole No. 497).

ATKINSON, R. C. The observing response in discrimination learning. *Journal of Experimental Psychology*, 1961, **62**, 253–262.

ATKINSON, R. C., & ESTES, W. K. Stimulus sampling theory. In R. D. Luce, R. R. Bush, & E. Galanter (Eds.), *Handbook of mathematical psychology*. Vol. II, New York: Wiley, 1963. Pp. 121–268.

BINDER, A. Effects of altered frequencies upon recognition responses. *Journal of Experimental Psychology*, 1963, **66**, 553–559.

BINDER, A., & FELDMAN, S. E. The effects of experimentally controlled experience upon recognition responses. *Psychological Monographs*, 1960, **74**(9, Whole No. 496).

BOURNE, L. E., JR., & RESTLE, F. Mathematical theory of concept identification. *Psychological Review*, 1959, **66**, 278–296.

BROADBENT, D. E. *Perception and communication*. New York: Pergamon Press, 1958.

ESTES, W. K. Of models and men. *American Psychologist*, 1957, **12**, 609–617.

ESTES, W. K. The statistical approach to learning theory. In S. Koch (Ed.), *Psychology: A study of a science*. Vol. II. New York: McGraw-Hill, 1959. Pp. 380–491. (a)

ESTES, W. K. Component and pattern models with Markovian interpretations. In R. R. Bush & W. K. Estes (Eds.), *Studies in mathematical learning theory*. Stanford, Calif.: Stanford University Press, 1959. Pp. 9–52. (b)

ESTES, W. K. Probability learning. In A. W. Melton (Ed.), *Categories of human learning (Proceedings of the Michigan-ONR conference on human learning)*. New York: Academic Press, 1964. Pp. 89–128.

ESTES, W. K., BURKE, C. J., ATKINSON, R. C., & FRANKMANN, J. P. Probabilistic discrimination learning. *Journal of Experimental Psychology*, 1957, **54**, 233–239.

ESTES, W. K., & HOPKINS, B. L. Acquisition and transfer in pattern-vs.-component discrimination learning. *Journal of Experimental Psychology*, 1961, **61**, 322–328.

ESTES, W. K., & STRAUGHAN, J. H. Analysis of a verbal conditioning situation in terms of statistical learning theory. *Journal of Experimental Psychology*, 1954, **47**, 225–234.

FELDMAN, S. E. Probabilistic hierarchies to ambiguous concept classes. *Journal of Experimental Psychology*, 1963, **65**, 240–247.

FRIEDMAN, M. P. Transfer effects and response strategies in pattern-versus-component discrimination learning. *Journal of Experimental Psychology*, 1966, **71**, 420–428.

FRIEDMAN, M. P., & GELFAND, H. Transfer effects in discrimination learning. *Journal of Mathematical Psychology*, 1964, **1**, 204–214.

LAWRENCE, D. H. Acquired distinctiveness of cues. II. Selective association in a constant stimulus situation. *Journal of Experimental Psychology*, 1950, **40**, 175–188.

Peterson, L. R. Prediction of response in verbal habit hierarchies. *Journal of Experimental Psychology*, 1956, **51**, 249–252.

RESTLE, F. A theory of discrimination learning. *Psychological Review*, 1955, **62**, 11–19.

SCHOEFFLER, M. S. Probability of response to compounds of discriminated stimuli. *Journal of Experimental Psychology*, 1954, **48**, 323–329.

TRABASSO, T. R., & BOWER, G. H. Component learning in the four-category concept problem. *Journal of Mathematical Psychology*, 1964, **1**, 143–169.

(Received May 16, 1966)

Vol. 80, No. 24 Whole No. 632

Psychological Monographs: General and Applied

EVOKED CORTICAL POTENTIALS IN RELATION TO CERTAIN ASPECTS OF VISUAL PERCEPTION[1]

CARROLL T. WHITE[2] AND ROBERT G. EASON

United States Navy Electronics Laboratory, San Diego

United States Navy Electronics Laboratory, San Diego State College

Evoked cortical responses were obtained in a number of studies dealing with various aspects of visual perception. On the basis of the variations noted in the complex response pattern under the different conditions it has been possible to identify certain components of that pattern as being related to specific aspects of the stimulus situation, such as intensity, color, and background level. In addition, the overall evoked response pattern appears to be directly related to phenomena encountered in the study of the perception of flickering stimuli.

The development of average response computers has been of great value to those involved in the study of human sensory mechanisms and perception, having made it possible to conduct both psychophysical or perceptual studies and neurophysiological studies with the same "intact" subjects (*S*s). This marks an important step toward the realization of Fechner's proposed "inner" psychophysics. The information to be gained by the use of this technique is of course of an entirely different nature than that being obtained by microelectrode techniques. Hopefully, it might relate more closely to integrative processes of the central nervous system and thus help bridge the gap between single-cell responses and subjective experience.

For the past several years our group has been studying human evoked cortical responses, primarily using visual stimuli. In order to better understand the nature of these responses, and to establish the range of conditions over which they might be useful as "objective" correlates of perceptual phenomena, a wide variety of situations have been investigated. On the basis of the results of these studies it has been possible to identify certain components of the evoked response pattern as being related to various aspects of the stimulus situation. In the present paper, data from a number of these studies are presented to illustrate the nature of the differences which have been found, and a tentative classification of the various components of the visually evoked response pattern is presented.

Effects of Variations in Stimulus Intensity and Background Level

This study was performed in order to provide a general idea of how the complex evoked response pattern changes under varying conditions of stimulus intensity and adaptation level. A "ganzfeld" stimulus was utilized, achieved by having half a table tennis ball attached over the eye stimulated, so that complexities which might be introduced by the presence of any marked contours in the field of vision would be avoided. In all the studies to be described in this report the stimulation was monocular (right eye), and the recording monopolar (between left occipital region and the left earlobe) unless stated otherwise.

The stimulus consisted of a single flash from a Grass photo-stimulator (P3) set at its highest output level (I-16). The light from the photo-stimulator entered *S*'s shielded cubicle through a filter holder which was mounted flush with a 2½-in. square hole cut in an opaque plastic "window" which was at *S*'s eye level. Immediately below this stimulus aperture

[1] This research was supported in part by National Science Foundation Grants GB-231 and GB-4067.

[2] The authors wish to thank John A. Hoke, M. Russell Harter, and Malcolm Lichtenstein for their assistance in various phases of this work.

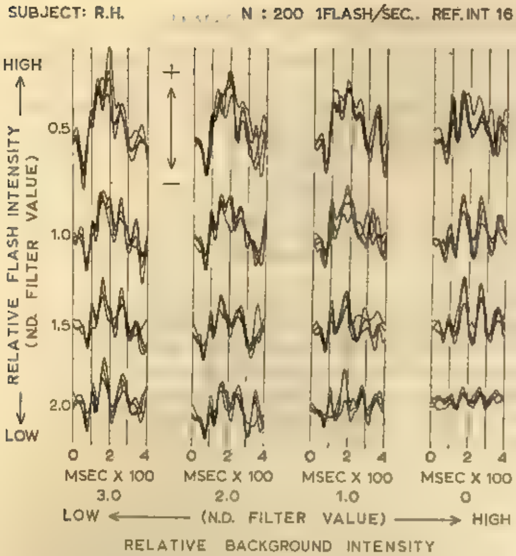

Fig. 1. Evoked cortical responses obtained under varying conditions of stimulus intensity and background level. Four replications of each condition, obtained on 4 consecutive days, are superimposed. White light stimulation. Negative downward.

there was a second similar arrangement through which the light from a small spotlight could be directed at the face of *S*, thus providing a variable background level on the "ganzfeld." The *S* was so positioned that his face was approximately 30 in. from the stimulus aperture.

In order to provide a suitable range of conditions the background level was set such that the stimulus was just detectable to *S* when the stimulus flash was reduced in intensity by means of a 2.0 log neutral density (ND) filter. Four intensity levels were used, separated by ½ log steps, and four background levels, separated by 1 log step. The stimulus flashes were presented at the rate of 1/sec. The responses to 200 such stimuli were summed by the computer of average transients (CAT) for each of the 16 conditions. This constituted a day's run for the *S*. Four such runs were conducted on each *S* on consecutive days so that an estimate of the degree of day-to-day variability in the response could be obtained. Two *Ss* were utilized for the complete series as described. A number of other *Ss* were tested under selected conditions so that an estimate of individual differences could be obtained.

The complete results for one *S* are shown in Figure 1. The four daily runs have been superimposed to show the degree of replicability obtained. The day-to-day variability shown in these records is in agreement with the reports of others who have studied this problem (e.g., Dustman & Beck, 1963). Individual differences in the overall waveform of the response were quite striking. The various components of the complex response pattern were present in all the individuals studied, but with differing relative amplitudes, thus creating markedly different waveforms for different *Ss*. This is also in line with the findings of other workers. The details of the response pattern can be seen more clearly in Figure 2, a single day's run for this same *S*.

Some general features are immediately obvious. The increase in the overall magnitude of the response as we go from the near-threshold condition at the lower right to the high-contrast condition at the upper left seems to agree quite well with the corresponding changes in perceived flash brightness. The same is true in general for the differences in overall amplitude within

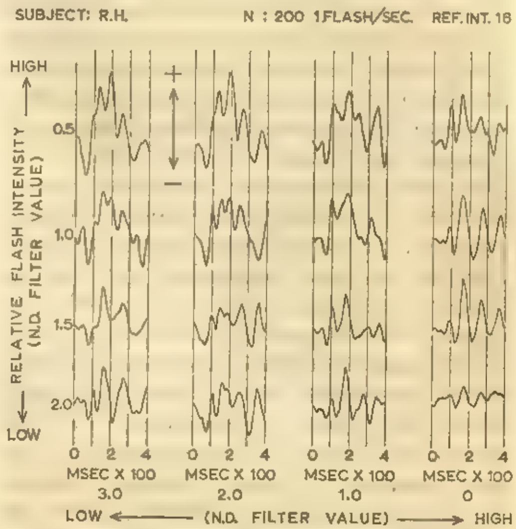

Fig. 2. Evoked cortical responses obtained under varying conditions of stimulus intensity and background level. One of the four replications shown in Figure 1.

the columns and within the rows. The very marked increase in amplitude that occurs in the immediate vicinity of the near-threshold condition should also be noted. We shall discuss this aspect in more detail in a later section.

In the preceding paragraph the term "overall" magnitude was used for a definite reason. A careful study of response patterns in Figure 2 shows that the various components that go to make up those patterns tend to behave quite differently as the experimental conditions are varied. To illustrate this point observe what happens to the three positive peaks occurring between 100 and 200 msec. after stimulation. These can be seen most clearly in the upper left quadrant of the figure. We shall refer to them as C_1, C_2, and C_3, in order of time of occurrence.

In the first column (3.0 log ND background), for example, observe what happens to the three peaks in question as the flash intensity is reduced. The change in C_3 (200 msec.) is most striking, it being almost eliminated completely when the stimulus intensity is decreased by one log unit. C_1 (100 msec.) is similarly affected by a decrease in flash intensity, but not to the same degree, there still being a well-defined response at the lowest stimulus intensity used. The amplitude of C_2, on the other hand, is relatively unaffected by the decrease in stimulus intensity. The main effect on C_2 appears to be a marked increase in peak latency as the stimulus is decreased. It should be noted at this time that the peak latencies of C_1 and C_3 do not change appreciably at a given background level, regardless of change in stimulus intensity.

Now observe what happens to our three positive peaks as we change background level, leaving stimulus intensity constant. The relative flash intensity of 1.5 (third row) illustrates these changes quite clearly. As the background level is increased we see that C_1 tends to increase up to a point and then to decrease as the background level is increased still more. The fate of C_2 and C_3 are even more interesting. C_2, the most outstanding feature of the response at the lowest background level, rapidly decreases in amplitude as the background is increased, being gone by the time the background has been increased by two log units. C_3, on the other hand, starts as a mere pip on the waveform pattern at the lowest background level but increases in amplitude very markedly as the background level is increased. Along with this increase in amplitude there is a significant decrease in the peak latency, amounting to from 30 to 40 msec. over the range of background intensities used.

We noted earlier that the amplitudes of Peaks C_1 and C_3 were directly related to stimulus intensity under the low-background situation. For the highest stimulus intensity used, therefore, these two are fairly large even at the lowest background level, and an increase in the background does not appear to bring about any great changes in their amplitudes (top row of figure). The decrease in the peak latency of C_3 with an increase in the background level does still occur. No such latency shift is noted for C_1, however. An examination of all the conditions shows this to be the case. The peak latency for C_1 did not change, even under conditions which produced marked changes in the peak latencies of C_2 and C_3. For the highest stimulus intensity condition, C_2 behaves as it did under the other conditions, decreasing in amplitude as the background level is raised, being entirely absent at the highest background level.

On the basis of the above observations some tentative hypotheses can be put forth regarding the nature of the three components in question. In the first place, it appears that C_2 is in some way related to *scotopic* visual activity. The fact that this component tends to be present only under the lower background conditions, and indeed is the outstanding positive component in the lower-left quadrant of the figure (representing both low-background and low-stimulus intensity) attests to this conclusion. On the other hand, both C_1 and C_3 appear to be related to *photopic* visual activity. The differences noted between the behavior of C_1 and C_3, however, lead to

the further conclusion that each is related to a different type of photopic activity.

In order to justify the conclusion that C_1 and C_3 probably relate to different photopic processes their differences should be reexamined. (*a*) In terms of peak latency, C_1 remained essentially constant over the entire range of stimulus intensities and background levels studied. The peak latency for C_3 decreased significantly under the higher background conditions. (*b*) In terms of amplitude, C_1 tended to vary directly with the stimulus intensity for any given background level. With C_3 this correspondence held only at the lower background levels. At the higher background levels some other factor seemed to be effective. Noting again the striking growth in amplitude of C_3 as a function of background intensity, especially for the lower stimulus flash intensities (and excluding the near-threshold condition), it seems very likely that this other factor might be related to the phenomenon of photosensitization or photic potentiation (Chang, 1959). The suggestion being made is that our C_3 seems to be related to a type of photopic activity that is greatly influenced by such a process.

Color Effects

The tentative identification of a scotopic component and two differing photopic components in the complex response pattern has led to the hope that color-specific response patterns might be identifiable. Figure 3 presents two examples of the kinds of results that give more reason for such a hope. Figure 3B is given here to illustrate the effect of a change in background level on the response pattern obtained from a different *S* (ML) than the one whose records have been discussed. In this particular situation the stimulus consisted of a combination of red and green (produced by Wratten filters 29 and 74). The result obtained under two conditions are shown superimposed. In one *S*'s room was darkened while in the other a medium-intensity background was projected onto the halved table tennis ball covering his eye. We would predict from the results of our first *S* (RH) that an increase in the background level should bring about a reduction in

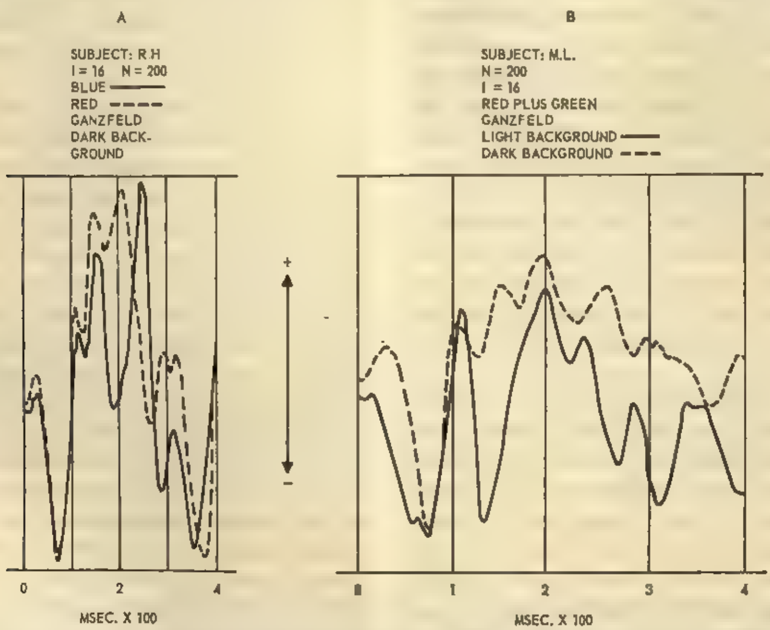

Fig. 3. A. Sample records illustrating the difference in *S* RH's responses evoked by red and blue stimulation. B. Sample records illustrating the effect produced by varying the background level. *S* ML. Note that the component occurring at around 150 msec. C_2 is most strongly affected.

amplitude in the region of time occupied by C_2 (about 130–160 msec.), whereas C_1 and C_3 might be accentuated. Such appears to be the case. At least the component relating to scotopic activity is common to both *S*s.

In Figure 3A we return to the first *S* (RH) to show his differing responses to red and blue light (Wratten numbers 29 and 49b). The highest flash intensity available was used (I-16 on the Grass photostimulator, with no neutral density filters) and the background was one log unit lower than the lowest used for the records shown in Figures 1 and 2. The difference in this *S*'s responses to red and blue is quite striking and obvious. Component C_3 appears to be entirely absent in the response to blue light. Instead, we have another positive component (C_4) that peaks about 35 msec. after C_3 should have appeared. The occurrence of a sizable C_1 under these conditions is further evidence that it relates to a different activity than does C_3.

Figure 4 presents another example of how the different components of the evoked response pattern may vary as a function of the stimulus color and how such differences can be enhanced by variations in the background level. This was also a ganzfeld situation. In the "ALL" condition the colored stimuli (Wratten numbers 29, 74, and 49b) were combined by allowing all three to strike the ganzfeld simultaneously. It should also be noted that this *S*'s response to blue is quite different than that of *S* RH (shown in Figure 3A). Component C_3 is definitely minimized in this case also, but C_4 is not such an outstanding feature. With this *S*, Component C_4 is more in evidence when the background illumination level was raised. The response patterns obtained for the "ALL" conditions were very similar to those obtained in other studies when this *S* was stimulated by white light.

The records shown in Figure 4 lend strong support to the conclusions arrived at earlier concerning the nature of C_1, C_2, and C_3. C_2 again seems to be related to scotopic activity, being sharply reduced in amplitude when the background illumination was raised. C_1 and C_3 again appear to be related to photopic activity, with C_1 being most responsive to green and C_3 being most responsive to red. These results indicate that this approach should be very fruitful for the study of responses to color and are quite in line with the results reported by Shipley, Jones, and Fry (1965), when allowances are made for differences in experimental techniques.

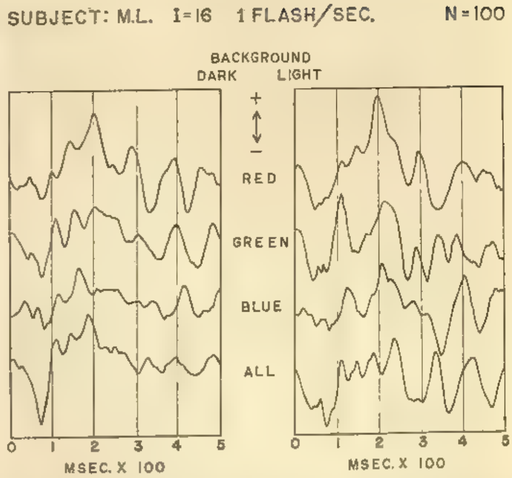

FIG. 4. Examples of *S* ML's responses to different stimulus colors. Note that the differences are enhanced by raising the background level.

EVOKED RESPONSE PATTERN AND PERCEIVED NUMBER

It has long been known that when one observes a flickering light source the perceived rate of flicker is not necessarily equal to the actual rate. As the result of an extensive series of studies on this topic, wherein trains of visual stimuli consisting of various numbers of flashes at various subfusional rates were presented to *S*s who reported the number of flashes perceived for each such train of flashes, the conclusion was reached that:

> the number of flashes reported by the subjects depended primarily on the time it took to present a stimulus sequence and not on the number of stimuli in that sequence. Perhaps the most impressive aspect of the data was the extreme reliability of the responses to certain number-rate sequences. For example, (for a certain level of background illumination) all the subjects always

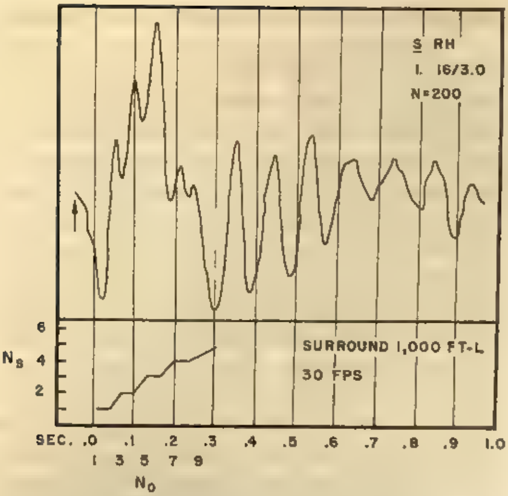

FIG. 5. An example of *S* RH's response to high intensity white light stimulation upon a low level background. Summation of 200 responses, with ganzfeld in use. Negative downward. Graph at bottom of figure represents the results of an earlier study in visual perception. N_s represents the number of "flashes" most often reported by *S*s when they were presented trains of various numbers of flashes (N_o) at the rate of 30 fps.

reported having seen 2 flashes whenever 5 flashes at 30 per second were presented to them, and they always reported having seen 3 flashes whenever 10 flashes at 30 per second were presented.... The results of this study convinced the authors that the limiting of the perceived number of flashes was due to some basic physiological process [White, 1963, p. 8].

The retina was at first suspected as the locus of this limiting process, but this was ruled out by studying the electroretinogram (ERG) obtained under high rates of photic stimulation. The retina was responding to every flash presented to it, so it was assumed that some more central process was limiting the number of flashes perceived. Further investigation led to the development of the "visual numerosity function," which showed the maximum number of flashes *S* could perceive as a function of the duration of stimulation. The visual numerosity function consists of two distinct segments: (*a*) from the onset of stimulation up to about 250–300 msec.; and (*b*) from about 250–300 msec. on. A more detailed description is as follows:

First, there is an initial fusion period, during which time the subject reports that he sees only a single flash; following this there is a short period when the function rises rapidly, the slopes indicating a rate of increase of perceived flashes of about 12–13/second. This rate is not maintained, however, but instead the function tends to level off about 200 milliseconds after the onset of stimulation. At about 250–300 milliseconds after onset, the second major segment of the numerosity function begins. As a result of all the studies done on this topic to date, there is good agreement to the fact that the slope of this portion of the function beyond 300 milliseconds indicates a rate of increase of perceived flashes of approximately 6–7 per second [White, 1963, p. 20].

The initial fusion period was found to vary with the background level, so it was classified as being related to a peripheral process (dark adaptation).

The general description of the numerosity function just given sounds very much like a description of the evoked response pattern itself. This pattern consists of two main parts, a complex transient response ending about 250 msec. after stimulus onset and a rhythmic afterdischarge which appears to start as the transient ends. Figure 5 shows such a pattern in rather pure form. This particular sample was obtained by summing the responses to 200 single flashes of white light. The time of the flash is indicated by the arrow. The response pattern has been shifted about 50 msec. to the left in relation to the time markers on the abscissa in order to adjust somewhat for the latency. The time values are meant to represent the time after the first neural activity evoked by the onset of stimulation has reached the cortex. This is a very rough approximation, but it will serve our present purpose. In the lower part of the figure, data from one of the series of studies on "temporal numerosity" are presented in graphical form. These particular results were obtained under conditions which have yielded the greatest perceived number of flashes as a function of flash-train duration (White & Cheatham, 1959). The values plotted are the modes. The ordinate (N_s) represents the number of flashes reported, while "N_o" is the number of flashes presented in a given flash train. N_o is plotted along the time

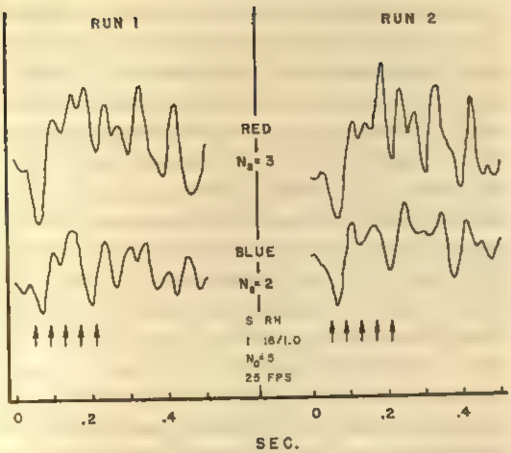

FIG. 6. Responses of *S* RH to trains of five red or blue flashes of light presented at 25 fps. N_s represents the number of "flashes" perceived by *S* under these two conditions. Run 1 and Run 2 were performed on consecutive days. Arrows indicating stimuli have been offset 50 msec. to allow (roughly) for latency. Ganzfeld condition used. Highest stimulus intensity available was used, presented upon a high-level background of white light. $N = 200$ for each record. Negative downward.

base to show how long it took to present a given train of flash stimuli at the repetition rate used in that study (30 fps.). For N_o's of one and two, the most frequent response was "one"; for N_o's of three and four, the most frequent response was "two"; for N_o's of five and six, it was "three;" for N_o's of seven and eight, it was "four;" while for an N_o of nine the responses were equally distributed between "four" and "five." When 10 flashes were presented at this rate the most frequent response made by *Ss* was "five."[8]

If these responses are considered in relation to the evoked response pattern in the upper portion of the figure a rather remarkable thing is seen—*the appearance in time of each successive perceptual unit seems to coincide with the occurrence of the successive components of the evoked response pattern.* Remembering that the evoked response pattern shown is produced by *single* flashes, this must mean that the onset of stimulation in some way initiates a process (or processes) which can have a marked influence on the perceptual response to any succeeding stimulation. The further implication is that the evoked response pattern should not be greatly changed by the presence of more than one flash in a sequence. Figures 6 and 7 show that such is indeed the case.

In Figure 6 we have examples of one of our *Ss'* response patterns evoked by trains of five flashes separated by 40-msec. flashes (25 fps.). Since this *S* had previously exhibited a markedly different evoked pattern in response to red and blue light (Figure 3A), this was also made a variable in this case. The two sets of response patterns, Run 1 and Run 2, represent replications of the experiment obtained on successive days. Each record represents the summation of 200 responses. The position of the five arrows representing the stimulus flashes have again been displaced in time in order to try to account somewhat for latency.

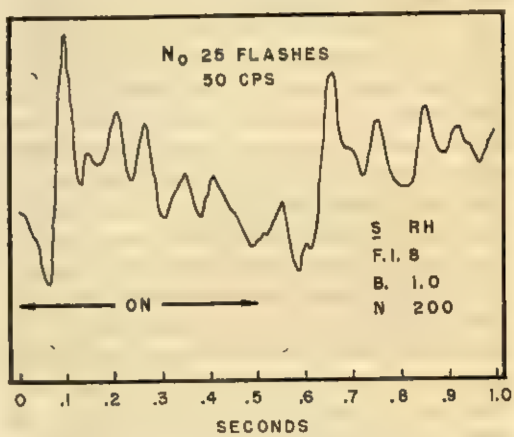

FIG. 7. Response pattern evoked by trains of 25 flashes of white light presented at rate of 50 fps. Ganzfeld condition, with relatively high-intensity stimulus (I-8 level of photo-stimulator) on high-level background. Summation of 200 responses. Note the marked off-effect. Negative downward. The *S* reported perceiving five or six "flashes" each time the stimulus train was presented.

[8] The difference in the number of flashes perceived in this case, as compared to the data quoted earlier, is explained by the fact that a much higher background level was used here. This minimized the duration of the initial fusion period. It should also be noted that the values given for the slope of the temporal numerosity function were based on the *mean* number of flashes perceived, and thus may be deceptively low.

As indicated in Figure 6, our *S* consistently reported having seen three flashes when the five red flashes were presented, and two flashes when the five blue flashes were presented. This difference seems to be related to the fact that one of the major components (C_3) is absent in the response pattern evoked by the blue stimuli. The results of this particular experiment tend to verify three things. First: the number of perceived flashes *S* reports in response to a train of flash stimuli is limited by the temporal characteristics of the cortical response pattern evoked by those stimuli. Second: the characteristics of the evoked cortical response pattern appears to be determined by the nature of the stimulus conditions at the time of onset. Third: this particular *S* exhibits reliably different evoked patterns for red and blue stimuli.

One last example of the relationship between the number of perceived flashes and the evoked cortical response pattern is shown in Figure 7. In this case, white light was used, again with the ganzfeld. The flash intensity was fairly high (setting number 8 on the photo-stimulator), upon a medium background. The evoked responses to 200 flash trains were summed. The stimulus trains consisted of 25 flashes, each flash separated by 20 msec. (50 fps.). Since in the various studies on "temporal numerosity" (the term which has been used to describe the perceived number phenomenon being discussed) the highest stimulus rate ever used was 30 fps., there were no perceptual data to compare with the evoked pattern. In this case it was decided to try to predict what the perceptual response would be on the basis of the evoked pattern. In Figure 7 it can be seen that there were six wave components during the period of time the intermittent light stimulus was being presented. Because of this it was predicted that the greatest number of flashes he would report having seen would be six. Other considerations, such as the possibility of an extended period of fusion immediately after onset of stimulation and the knowledge that the rate at which additional perceptual units are added to a perceived sequence decreases sharply after a duration of 300 msec., led to the conclusion that he would also report having seen only five flashes at least part of the time. All this time *S* remained in his shielded room unaware of the predictions which were made. Upon being asked to report the number of flashes perceived after each flash train was presented he began by reporting "six." With continued repetitions he began to report seeing only "five" part of the time. At no time did he report having perceived anything other than "five" or "six." Later, a number of other of our personnel were presented with this same stimulus train. All agreed that there appeared to be five or six flashes in the sequence.

As the result of the extensive studies on "temporal numerosity" which were performed it was concluded that the onset of stimulation triggers some central process (or processes) which interacts with afferent neural activity in such a way to limit the rate at which perceptual units can be added. The studies on evoked response patterns have shown that the onset of stimulation does indeed trigger some central cyclic processes, whose temporal characteristics are very much like those of the hypothetical processes. Both the perceptual data and the evoked response patterns show an important change in character at a point about 250 msec. after onset. This marks the end of the initial complex pattern and the beginning of the rhythmic aftereffect. In the perceptual data this marks the point where the rate at which perceived flashes are added to a sequence changes from about 12/sec. to about 6/sec. During this first segment there seems to be a one-to-one relationship between the components of the response pattern and number of flashes which can be preceived. During the second segment there seems to be a two-to-one relationship between the cyclic brain processes and the perceived events. (The basic frequency of the rhythmic aftereffect appears to be equal to *S*'s alpha rhythm.) Thus there definitely appears to be a close relationship between the temporal numerosity phenomena and the evoked cortical response pattern. The exact nature of this

correspondence and its functional significance, if any, are not clear. It is tempting, however, to view this relationship in terms of the cyberneticists' "scanning" mechanism (Wiener, 1948) and the concept of the "psychological moment" (Stroud, 1955; White, 1963). It is also of interest to note that the duration of the initial phase of the evoked response pattern (about 250–300 msec.) corresponds to a duration that has been found to be critical in studies dealing with complex visual discrimination, especially those associated with contour processes and identifications (Kolers, 1964; Schlosberg, 1965). This appears to give more substance to the concept, inherent in much of the previous discussion, that the components of the evoked response pattern are related in some way to the various aspects of the informational processing of visual stimuli.

EVOKED RESPONSE AND PERCEIVED BRIGHTNESS

In this discussion of the evoked response patterns shown in Figure 1 it was pointed out that the amplitude of some of the components seemed to vary directly with the intensity of the light stimulus. This is especially true of the component referred to as C_1 (peaking at about 100 msec. after onset of stimulation). It can be seen that for any given background level this is the case. Component C_3 (peaking at about 200 msec. after onset) seems to be useful in this regard only when the background level is relatively low.

A study was performed in order to determine how well the amplitude of these components would correlate with perceived brightness in a very marginal situation. Stimuli consisting of pairs of identical flashes separated by 9, 16, or 25 msec. were presented to a group of *S*s. All of these flash pairs were perceived as single flashes by *S*s. In order to establish the relative brightness of these three stimuli a temporal forced-choice procedure was carried out, the results of which showed that the perceived brightness decreased as the interflash interval increased. The second phase of the study consisted of obtaining evoked response patterns for the two smaller inter-

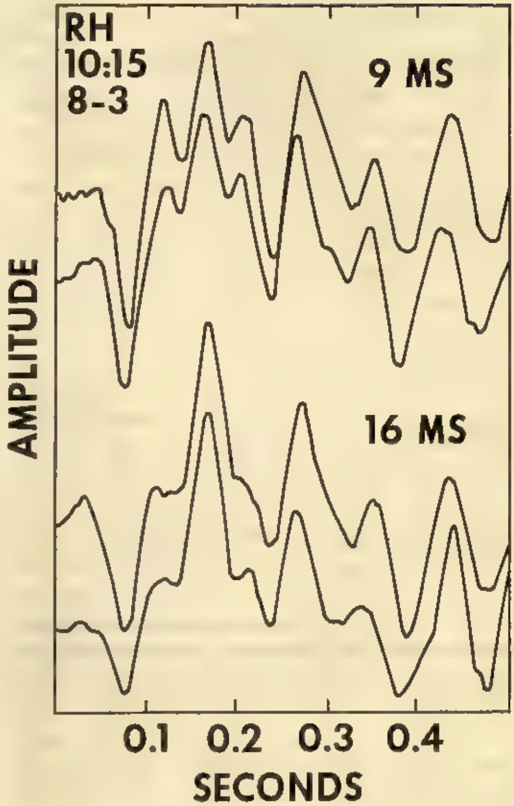

FIG. 8. Evoked potentials obtained in response to fused flash pairs having interflash intervals of 9 and 16 msec. Onset at start of trace, each trace representing the summation of 100 flash pairs in one channel of the computer. All four records obtained during a single session, in counterbalanced order. Negative downward.

flash conditions (9 and 16 msec.). It was found that the amplitudes of the critical components (C_1 and C_3) were significantly different for the two conditions. Figure 8 illustrates the type of results obtained. (Each condition was replicated twice in this study in order to check on the reliability of the results.)

It is interesting to compare the response patterns in Figure 8 with those in Figure 2, which were obtained earlier from this same *S*. The differences noted between the responses to the two conditions in Figure 8 are seen to be very much like those shown in the first column of Figure 2 for relative flash intensities of 1.0 and 1.5.

This particular study has been reported

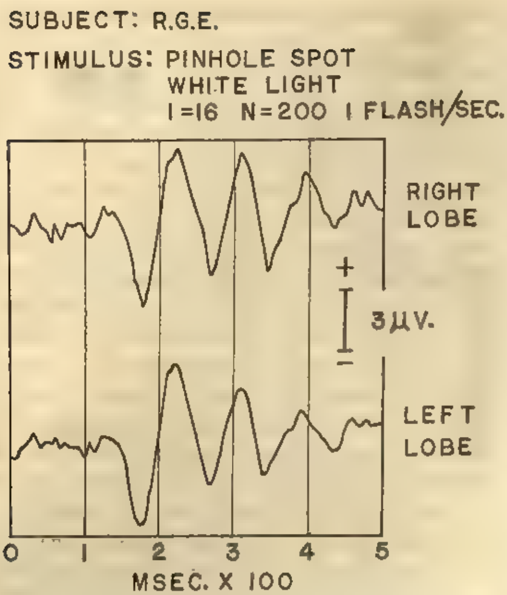

Fig. 9. Evoked response patterns recorded simultaneously from scalp over both occipital lobes. High intensity pinhole source, fixated foveally, with medium background level. Summation of 200 responses.

in greater detail elsewhere (Bartlett & White, 1965).

Evoked Responses to Minimal Visual Stimuli

The studies described so far have all been concerned with the responses elicited by full-field stimulation. If there are basic differences between the responses elicited by stimulation of foveal and peripheral regions of the retina, as is indeed the case,[4] they would not be revealed by such procedures. For this reason, among others, it was decided to determine the nature of the response to minimal stimulation, both in terms of the physical size of the stimulus and its intensity.

One important result of our work along this line was the general finding that if a given stimulus situation was perceived by *S* an evoked response could also be obtained. This was noted earlier in regard to Figure 1 and Figure 2 for the base condition (near-threshold) in the lower right corner. The present case was a more dramatic illustration of this principle, however. In trying to discover how small a visual stimulus we could use and still obtain an evoked response it was found that there seemed to be no effective lower limit—if the stimulus were visible a response would be elicited. So, for our minimal visual condition we used what was quite literally a pinhole source. The face of the photo-stimulator was masked by black electrician's tape, in the center of which a very minute hole was produced by the point of a needle. A piece of tracing paper was placed flush with the masking tape, both to provide a diffuse light source and a white surround. The location of the stimulus hole was indicated by a black circle about 2 mm. in diameter. The background light was raised to a level such that the fixation circle could be seen clearly.

A sample response pattern obtained under the conditions just described is presented in Figure 9. Here the highest intensity produced by the stimulator was being utilized and was perceived by *S* as an intense spot of light. It can be seen that the form of the response is very similar to those shown in Figure 2 in the right-hand column, where the full-field stimulus was rather weak and the background level was high.

Since we could obtain an evoked response to this very minute stimulus it was decided to see how the nature of the response would vary as a function of various stimulus parameters. Of particular interest was the effect of stimulus duration. In one study, trains of flashes (with an interflash interval of 10 msec.) were used to provide varying durations of stimulation. The lowest intensity level produced by the photo-stimulator (I-1) was used, and the intensity of the background was adjusted so that a single flash was approximately at *S*'s threshold. As additional flashes were added to the trains the perceived brightness and the apparent size of the stimulus light increased markedly. The evoked response paralleled this increase, being barely detectable for the one-flash condition and

[4] An extensive study dealing with variations in the evoked response as a function of the location of the stimulus in the visual field has been performed by Eason and White (in preparation). In addition to such position effects, marked lobe dominance effects were indicated.

increasing to a maximum amplitude for the four-flash condition, representing a total duration of stimulation of 30 msec. It was assumed that this was the critical duration for the conditions of the study. A comparison study was carried out using a glow modulator tube as the light source. Here the duration of a continuous light source could be varied. The results were similar to those obtained with the fused flash trains. A detailed report on these two studies is being prepared.

DISCUSSION

It has been shown that the evoked cortical response, as obtained with a device such as the computer of average transients, can be of great value in the study of a wide variety of visual problems. Stimuli ranging from minute point-sources up to the ganzfeld can all be utilized, as can situations involving thresholds, brightness perception, and color vision. Certain studies dealing with the temporal characteristics of the visual process have also been successfully carried out, both by ourselves (as described above) and by other workers (Donchin, Wicke, & Lindsley, 1963). Preliminary studies comparing the responses evoked by ganzfeld stimulation with those evoked by structured fields have suggested the possibility that certain aspects of form perception could also be approached by this technique.

The types of studies mentioned in the preceding paragraph all deal with the physical aspects of the stimulus situation. A number of previously published reports by various workers have demonstrated the marked effect of subjective factors on the evoked response patterns. These have been referred to by such terms as "attention," "vigilance," "level of activation," and "meaningfulness of the stimuli" (e.g., Chapman & Bragdon, 1964; Eason, Aiken, White, & Lichtenstein, 1964; Spong, Haider, & Lindsley, 1965). All such factors probably contribute to the intrasubject variability of response which is found by all workers in this field. The well-established habituation of response to repetitive stimulation, undoubtedly related to the subjective factors listed, is also of importance in this regard.

The changes in evoked response patterns related to the subjective factors listed are of great interest and value, but when one is trying to relate those patterns to psychophysical phenomena such variability is most troublesome. Experience has shown that one must expect such variability in any study being contemplated and do everything possible in the way of experimental design to minimize the effect. Short runs of any one condition, frequent opportunities for *S* to leave the experimental area, and counterbalancing of the various conditions over time are all essential. The fact that there is marked intersubject and intrasubject variability must be considered to be a blessing instead of a curse, since it attests to the sensitivity of the technique. Such sensitivity places greater demands on the ingenuity of the workers using the technique if they hope to derive the full benefit from it, but the end result should be well worth the added effort.

The examples of evoked response patterns which have been presented show that there are marked differences in the nature of the response under various conditions of stimulation. At one extreme is the very simple oscillatory pattern produced by the pinpoint of light striking the fovea; at the other is the complex pattern produced by high-level ganzfeld stimulation. On the basis of the changes noted in the response pattern under various conditions (color, intensity, background level, and retinal locus of stimulation) the conclusion has been reached that there are a number of component responses, each related to some aspect of the stimulus situation. It is further concluded that the evoked response pattern produced by the high-level ganzfeld stimulation is a composite of these various responses. In Figure 10 a sample response pattern to ganzfeld stimulation is presented, along with a tentative breakdown of the component elements.

Process I appears to be related to the degree of scotopic activity. Its amplitude varies with the relative intensity of the stimulus flash—becoming greater as the background is reduced with a given flash intensity and also as the flash intensity is increased with a given background level (Figures 1 and 2).

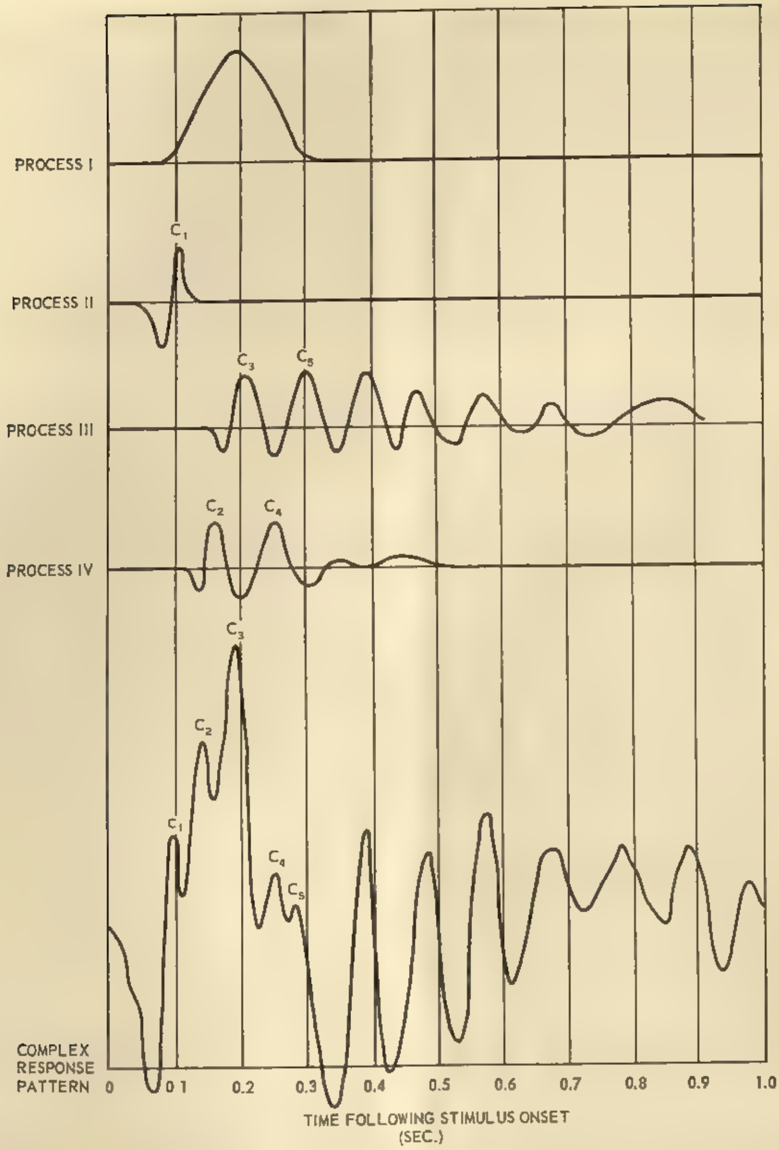

Fig. 10. Tentative classification of processes and components making up the complex response pattern evoked by high intensity white light stimulation in the ganzfeld condition. Sample record shown is the same as that in Figure 5.

Process II is assumed to be related to photopic activity. Data from the color studies and the brightness discrimination study (both described earlier) indicate its probable relation to a green response and/or a photopic brightness discrimination mechanism. Its peak latency is relatively constant under all conditions of stimulus intensity and background level.

Process III is assumed to be related to some aspect of photopic activity. This assumption is based on the color studies, where Peak C_3 was so evidently related to a red response (Figures 3, 4, and 6), and on the pinhole stimulus situation (Figure 9) wherein only the fovea was stimulated. This process is differentiated from the other presumably photopic response, Process II, by the fact that its latency varied considerably as a function of stimulus intensity and background level. Its latency is seen to decrease as the *total* amount of

light in the stimulus situation increased, including both the stimulus flash and the background level (see Figure 2).

Process IV has tentatively been classified as being related to a blue color and/or a scotopic response mechanism. The reasons for this were pointed out earlier, and are clearly shown in Figures 3 and 4. The positive peak, C_2, is seen to be very sensitive to background level, its amplitude decreasing markedly as background intensity is increased. The probable relationship of Peak C_4 to a blue response mechanism is shown most strikingly for *S* RH in Figure 3.[5] It has been found for this *S*, with blue light stimulation, that Peak C_2 is more sensitive to changes in background level and Peak C_4 is more sensitive to changes in flash intensity. This would also suggest that C_4 might be related to a photopic response to blue light.

One final comment should be made regarding the possible significance of the various components of the complex evoked response pattern. A preliminary study in which responses to ganzfeld stimulation were compared to those evoked by a structured visual field indicated that the later components, around 250–300 msec. (C_4 and C_5), were of much greater relative amplitude in the structured field situation. If further investigation verified this point it is believed that this may be relevant to perceptual studies on the "serial processing of visual information" such as those of Kolers (1964), who found that approximately ⅓ sec. was necessary for the assimilation of contour information.

The form of the evoked response patterns obtained is dependent on a number of factors regarding experimental technique, so it is difficult to compare one's results with those of other workers unless the identical procedures were followed. Electrode placement, the use of monopolar or bipolar electrodes, and the time between stimulus presentations are three such factors which can lead to a wide divergence in the results obtained. In all the situations reported here a monopolar occipital electrode was utilized, and the stimuli were always presented about 1 sec. apart. Other variables of importance, such as whether a ganzfeld stimulus or a restricted field was used, are indicated in the various sections.

The portion of the evoked response patterns which we have been concerned with in this paper corresponds to that which Cigánek (1961) has termed the "secondary response" and Spreng and Keidel (1963) refer to as the "medium components." These authors agree that the characteristics of these components indicate that they are related to activity of the nonspecific, perhaps diffuse afferent pathways. These characteristics are the long latencies involved, the fact that these components can be recorded from broad areas of the scalp, and the fact that they are elicited by stimulation of various sense modalities. The point of interest here, however, is that correlations with specific aspects of stimulation within a given modality can be obtained if inputs from the other modalities are precluded.

Even though certain of the components occurring during the period following stimulation in question (roughly 80–300 msec.) may be evoked by the various modalities, it is still quite possible that some of the other components might be related to a specific type of input. For example, the prominent "medium components" evoked by acoustical stimulation, as presented by Spreng and Keidel (1963), appear to be most closely related to those components we have tentatively identified as being related to *photopic* activity in our work with visual stimulation (i.e., the components designated as C_1 and C_3). It seems quite possible that the unique dual nature of the visual system, wherein the scotopic-related neural activity occurs later than that of the photopic-related activity, might well give rise to evoked components that

[5] It should be pointed out that the response patterns shown for RH in that figure were taken from parametric studies which were carried out for various colors of the stimulus flash. For each stimulus color used, flash intensity and background level were varied as they were for the white stimulus flashes in the first study described (Figures 1 and 2). We are not planning to publish the more detailed report on color responses until we have a chance to run complete parametric studies on more subjects, so that a more meaningful comparison of individual differences can be made.

are modality specific. The component we have designated as C_2, which appears to be related to scotopic activity, is suggestive in this regard. In other words, it might be well to consider the photopic and scotopic visual systems as two separate submodalities in view of the different time courses of their neural activity following stimulation.

There is one rather major way in which our tentative description of the events following the onset of stimulation differs from those of the other workers referred to. The striking change in the character of the complex evoked response pattern which occurs around 250–300 msec. after the onset of stimulation is usually described as marking the end of the "secondary response" or "middle components" and also the beginning of the "rhythmic" or "oscillatory" afterdischarge. There is the definite implication that this afterdischarge is of a different nature than that occurring during the "secondary response" in such a description. In our studies utilizing as purely a photopic stimulus as possible (a pinpoint source of light, fixated foveally, with a light-adapted eye) it was found that the oscillations clearly started between 100 and 200 msec. after stimulus onset and continued on into the time domain of the so-called "afterdischarge" (Figure 9). Therefore we interpret the oscillatory activity occurring after 250–300 msec. following onset of stimulation as being a continuation of a process which was initiated during the time of the "secondary response."

The marked change in the character of the waveform in the region of 250–300 msec. is believed to be related to the dying out of other processes, which do exist only during the period of the "secondary response." These processes are the ones designated as I and IV in Figure 10. The termination of these processes would leave Process III, in relatively pure form, as the oscillatory afterdischarge.

REFERENCES

BARTLETT, N. R., & WHITE, C. T. Evoked potentials and correlated judgments of brightness as functions of interflash intervals. *Science*, 1965, **148**, 980–981.

CHANG, H. T. The evoked potentials. In H. W. Magoun, (Ed.), *Handbook of physiology, section I: Vol. 1 Neurophysiology*. Washington, D.C.: American Physiological Society, 1959.

CHAPMAN, R. M., & BRAGDON, H. R. Evoked responses to numerical and non-numerical visual stimuli while problem solving. *Nature*, 1964, **203**, 1155–1157.

CIGÁNEK, L. The EEG response (evoked potential) to light stimulus in man. *Electroencephalography and Clinical Neurophysiology*, 1961, **13**, 165–172.

DONCHIN, E., WICKE, J. D., & LINDSLEY, D. B. Cortical evoked potentials and perception of paired flashes. *Science*, 1963, **141**, 1285–1286.

DUSTMAN, R. E., & BECK, E. C. Long-term stability of visually evoked potentials in man. *Science*, 1963, **142**, 1480–1481.

EASON, R. G., AIKEN, L. R., WHITE, C. T., & LICHTENSTEIN, M. Activation and behavior: II. Visually evoked cortical potentials in man as indicants of activation level. *Perceptual and Motor Skills*, 1964, **19**, 875–895, Monograph Supplement 3-V19.

KOLERS, P. A. Apparent movement of a Necker cube. *American Journal of Psychology*, 1964, **77**(2), 220–230.

SCHLOSBERG, H. Time relations in serial visual perception. *Canadian Journal of Psychology*, 1965, **6a**, 161–172.

SHIPLEY, T., JONES, R. W., & FRY, A. Evoked visual potentials and human color vision. *Science*, 1965, **150**, 1162–1164.

SPONG, P., HAIDER, M., & LINDSLEY, D. B. Selective attentiveness and cortical evoked responses to visual and auditory stimuli. *Science*, 1965, **148**, 395–397.

SPRENG, M., & KEIDEL, W. D. Neue Möglichkeiten der Untersuchung menschlicher Informationsverarbeitung. *Kybernetik*, 1963, **1**, 243–249.

STROUD, J. M. The fine structure of psychological time. In H. Quastler (Ed.), *Information theory in psychology*. Glencoe, Ill.: Free Press, 1955. Pp. 174–207.

WHITE, C. T. Temporal numerosity and the psychological unit of duration. *Psychological Monographs*, 1963, **77** (12, Whole No. 575).

WHITE, C. T., & CHEATHAM, P. G. Temporal numerosity: IV. A comparison of the major senses. *Journal of Experimental Psychology*, 1959, **58**, 441–444.

WIENER, N. *Cybernetics*. New York: Wiley, 1948.

(Received May 27, 1966)

Vol. 80, No. 24 Whole No. 633, 1966

Psychological Monographs: General and Applied

Duplicate

HETEROMODAL EFFECTS UPON VISUAL THRESHOLDS[1]

EDWARD T. DAVIS

Veterans Administration Hospital, Bedford, Massachusetts

Processes underlying the transmission and coordination of 2 different kinds of sensory excitation were studied. A neurological model accounting for specific heteromodal effects was proposed. The method involved the determination of visual thresholds in normal and brain-injured Ss while they were being subjected to an auxiliary aural stimulus of moderately loud intensity. The results demonstrated group differences in the effect sound has on visual thresholds and provided information on the diminishing effectiveness of a constant auxiliary stimulus when it is maintained for a period of several minutes. The findings were reviewed in the light of past and present theoretical explanations and related to a brain model which accounts for both facilitative and inhibitory effects of auxiliary stimulations.

The mechanisms by which all kinds of everyday sensory excitations are somehow compounded into a meaningful experience for an individual must certainly be an action of a most complex order. Although we have little knowledge of the processes mediating the final consummation of experience, there can be little doubt that there are definite physiological prerequisites for this phenomenon.

One approach to the problem just mentioned has been the study of very basic intersensory effects such as the influence a sound has on an absolute visual threshold. As a means of making intelligible the results of such studies, most investigators have offered theoretical explanations which involve the transmission and coordination of neural excitation in the central nervous system.

In an attempt to avoid complexities that are apt to occur in any perceptual study, experimenters have tried to eliminate any overlay of meaning from the stimuli they have used. Attempts have been made to keep the subjects as objective as possible, and thus tasks involving a minimum of interpretation are the most useful. Usually a small white patch of light and a reasonably pure tone of a given frequency are used in such studies.

The study to be outlined here is an attempt to clarify elementary heteromodal relationships. The objectives may be briefly stated as follows:

1. To extend the examination of heteromodal processes by considering temporal and intensity factors in the effects of an auditory stimulus on a visual threshold.

2. To study subjects suffering from severe brain injury in the hope of supporting or opening to question some aspects of neural explanations.

3. To develop a theoretical model which would allow a parsimonious explanation of these extended intersensory relationships.

The exact processes underlying the transmission and coordination of excitation in the cortex has intrigued both psychologists and neurophysiologists. Köhler and Wallach (1944) and Lashley, Chow, and Semmes (1951) have used behavioral criteria as a means of exploring electrical energy transmission in intracortical processes and Chang as early as 1952 ventured to suggest a type of neural process that could account for the influence of auditory stimulation on a visual pathway. Jung (1961) and Jung, Kornhuber, and Da Fonseca (1963) have summarized a wealth of infor-

[1] Based in part on a doctoral dissertation submitted to the Department of Social Relations at Harvard University. The author wishes to express his appreciation to G. S. Klein for his guidance and encouragement. Additional thanks go to W. S. Verplanck and G. A. Miller whose advice and technical knowledge were essential in determining the method of measuring visual thresholds in this study.

mation, based on the electrocortical studies of many investigators. They have related neural functioning to a variety of subjective phenomena which more often are thought to lie within the province of the psychologist.

The theoretical background of the present study was influenced by Hebb's (Hebb, 1949) attempt to apply neurophysiological concepts to the explanation of behavior. It seemed reasonable to assume, as Hebb did, that today's neurophysiology could provide facts and theories that would allow the construction of a testable hypothesis for the examination of intersensory effects. Many of the sensory pathways can be shown to be anatomically proximate at some point in the brain; and recent investigators, both psychologists and neurophysiologists, have theorized that in the action following heteromodal stimulation some type of neural communication must exist to account for the results obtained by accessory stimulation.

Present explanations suggest that the neurons of the tested modality are increased in excitability because of an anatomical proximity to neurons of the accessory modality that are firing at the same time. The increased excitability is based on the theory that the local potential of neurons in the tested modality is raised by the electrical activity of the neurons firing in response to the accessory stimulus. In this way the threshold is lowered. In such a theory, a neuron that is capable of responding to the excitation of at least two modalities is an essential element.

Early experiments demonstrating the convergence of different sense modalities on neurons of the reticular formation (Amassian & DeVito, 1954; Baumgarten, Von Mollica, & Moruzzi, 1953; Scheibel, Scheibel, Mollica, & Moruzzi, 1955) have led to the examination of many other polysensory areas including the cortex. More specifically, single unit analysis has shown that the same cell may be fired by afferent volleys from different sensory pathways (Amassian, 1954; Segundo & Machne, 1956).

From a psychological point of view, however, these studies raise an important question. Once such a polysensory element is fired, to which sensory experience does such a mutually recruitable neuron contribute its effect?

Hebb (1949) has suggested that time relationships are of extreme importance in neural organization. Single neural units become functionally integrated cell assemblies through simultaneity and ordering of firing. Thus a neuron may contribute to one "phase sequence" or another, depending upon the time at which it fires. Segundo (Segundo & Machne, 1956) has suggested that a convergence of this type upon a single neural element does not necessarily mean complete loss of that level of the capacity to discriminate between different sensory stimuli. He points out that the temporal pattern of response of that unit to each stimulus may be very different. In this manner information may be preserved on the basis of temporal criterion even in the presence of spatial convergence. Such a possibility would allow the neuron that responds to two sensory excitations to contribute its effect to first one and then the other sense modality. Its selectivity would depend upon the relative rate at which the two modalities bombarded it with excitation. That is, the modality that bombarded it with the greater frequency would have the better chance of finding such a mutually recruitable neuron in a nonrefractory state and firing it. Since frequency of neuron discharge is related to intensity of stimulation, the sensory receptor receiving the greater amount of excitation would "steal" or "capture" the disputed neuron. It does not mean, however, that subliminal excitation arriving from the accessory stimulus might not lower the threshold of a neuron and thus facilitate its firing in the primary system by spatial or temporal summation.

The existence of many such neurons varying in threshold of excitability is the key postulate in the theoretical organization of this thesis. Such a hypothetical neuronal scheme has been postulated by Jung (Jung et al., 1963). In general the connections are inferred from the results of neu-

ronal recordings from the cat's cortex. He has conjectured that in as much as reticular neurons already receive the convergence of several sensory modalities, a subcortical component of multimodality input is very probable. However, he also includes the possibility of intercortical connections.

The postulates used in this thesis are necessarily psychological in nature since the data collected are gleaned from behavioral events. They are stated with no intention of testing neurological concepts, but the underlying reasoning for these assumptions is based on the theorizing of previous investigators and is extended and summarized by the use of a hypothetical brain model.

In order to trace the development of the thinking in this paper, essential characteristics of the brain model are described and summarized immediately prior to the specification of the formal postulates associated with this study. Both types of explanation are developed simultaneously with the intention of making the two levels of reasoning clearly distinguishable.

Figure 1 depicts a neurological model offered as a schematic analogue of the processes under investigation. It is borrowed from the thinking of Jung (Jung et al., 1963) and is presented with the expectation that it is capable of generating heuristic possibilities.

As Fessard (1961) has pointed out, neural models are simplified neural circuits of special design and it is unsound to infer close resemblances in internal organization and the performance of other operations known to occur in nature without substantial support for such reasoning. Jung (Jung et al., 1963), however, believes that such a caution does not preclude the expediency of developing common lines of reasoning in the reduction of problems that require both neurophysiological and psychological knowledge. Licklider (1961) has suggested that psychophysical models offer the possibility of bringing together in productive interaction, facts and findings from a variety of sources. To support this view he has proposed a model drawn from a variety of disciplines which accounts for the analgesic effect music and noise have on pain thresholds.

Since there are, as yet, no definite neurophysiological observations underlying effects found in this study, the explanation offered is admittedly speculative. Although speculation may be hazardous it fulfills an important purpose when it offers perspective to a field of inquiry or when it stimulates research designed to replace speculation by factual demonstration. It is hoped that the following model fulfills both of these criteria.

POSTULATES AND THE ASSOCIATED BRAIN MODEL

The model represents a neural network assumed to exist in the human brain. Neurons I and II are neurons in primary pathways serving two different senses. The neural chains linking Neuron I and Neuron X, and Neuron II and Neuron X are functioning cerebral parts of their respective sensory modalities (I and II). Neuron X is a cerebral neuron. It is assumed that the network represented by Figure 1 exists in large numbers in the human brain. The functional properties of this network include the following elements: With each sensory modality there are a number of cortical neurons which may be fired by more than one kind of modal excitation and which contribute to the intensity of experience of the modality by which they are fired. The frequency of neural discharge in a modal network is positively related to the strength of the sensory stimulus.

Postulate 1. Any sense modality has a process P which may be stimulated by stimuli of that modality or by stimuli from other modalities (e.g., P_v is a function of the visual stimulus I_v, and/or the

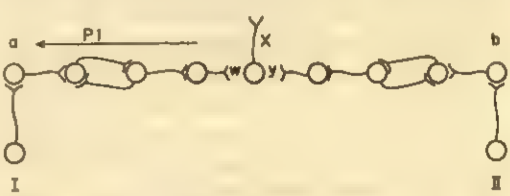

FIG. 1. Neurological brain model.

intensity of other stimuli such as I_h, and aural stimulus.

It has already been suggested that a mutually recruitable neuron (or mutually recruitable neurons) is selected and fired by one modality or the other. Implicit in this description is the assumption that temporal relationships allow neural threshold responses to contribute to only one modality at a time. Thus, if a group of neurons which were originally a part of the visual process P_v are more heavily bombarded by excitation from another modality they shift their effect to the latter system. That is, once fired a neuron contributes its effect to either the visual or auditory processes but not to both. The relationships of such an assumption are more generally stated in the following postulate form:

Postulate 2. If P is aroused to a threshold of contribution by I_h it is no longer aroused to a threshold of contribution by I_v and no longer functions as a part of the I_v system.

The process by which mutually recruitable neurons shift their effect has already been described as positively related to the frequency which a given modality excites them. That is, the probability of a given mutually recruitable neuron (X) being fired by a given modality (P 1) is a direct function of the ratio of frequency of neural discharge at that model synapse with X (w) to the sum of all modal discharge frequencies synapsing with X (w plus y) that is,

$$P\,1X = f\,\frac{\text{frequency M 1}}{\text{frequencies M 1} + \text{M 2} \cdots \text{Mn}}$$

The relationship is postulated in the formal modal as follows:

Postulate 3. P_v bears a relationship to visual and auditory stimuli at levels of threshold contributions such that:

$$P_v = (f)\,\frac{\text{intensity of visual stimulus}}{I_v + I_h}$$

$$= (f)\,\frac{I_v}{I_v + I_h}$$

It has been demonstrated that neural tissue becomes fatigued under continuous stimulation; that is, the frequency of firing of a given fiber drops even though the strength of the stimulus remains the same: that is, with a constant stimulus, the frequency of neural discharge in its modal network does not remain constant but reduces to approach a stable lower level as the stimulus is maintained. This effect is accounted for in the following postulate:

Postulate 4. The value of I_v and I_h is a function of the duration of the appropriate stimulus, S_v and S_h which has produced them. The relationship is such that the intensity decreases with time since the application of the stimulus.

With the experimental procedure used in this study, the visual stimulus is not as constant as the auditory tone used. The visual stimulus fluctuates in intensity and is often interrupted by the involuntary eye blinking of the subject. Since neurons regain their excitability after continuous stimulation within a short recovery period, this relationship is summarized thus: The rate of decline of the frequency of neural discharge in a modal pathway is positively related to the constancy of the stimulus impinging upon this avenue of excitation. It is formally accounted for in Postulate 5:

Postulate 5. The value of I_v for a given S_v tends to maintain its maximal value if a recovery period exists between the cessation of the last application of S_v and the current application of S_v.

Since the data in this study are psychological data, it is necessary to postulate a relationship between the assumptions just listed and the behavior tested. Neurological evidence demonstrates that the intensity of experience of a sensory stimulation is positively related to the number of cortical neurons fired by that stimulus in its modal area. From this the formal statement becomes Postulate 6:

Postulate 6. The reported visual threshold is proportional to the reciprocal of P (i.e., the larger P_v the lower the threshold T_v).

A consequence of the characteristics ascribed to the brain model allows the following prediction: A mutually recruita-

ble neuron will tend to be captured by the modality bombarding it with the greatest frequency of excitation.

The formal reasoning leading to the first theorem rests on the formal postulates. Since we have assumed that P_v can be measured by a visual threshold (Postulate 6), certain effects of a moderately loud auditory stimulus on a visual threshold can be deduced.

Postulate 1 states that P can be aroused by excitations from the auditory and/or visual system. Postulate 2 states that a contribution effect of P cannot serve as a part of the visual and auditory systems at the same time. Since, by Postulate 3, P is recruited in a direct relation to the intensity of a given stimulus, then Theorem 1 follows:

Theorem 1. A moderately strong auditory stimulus will initially raise the visual threshold of normal subjects.

Again, since excitation resulting from stimulation decreases with time (Postulate 4) it is only necessary to assume the existence of a recovery period for the visual pathway in order to predict a recovery of the visual threshold with time (Postulate 5). Such a recovery period for the visual pathway could be produced by a fluctuating target light and the eye blinking associated with the fixation of a visual target. The auditory pathway would have no such respite. Thus, the more frequently excited and more constantly excited modal pathway will fatigue earlier.

With the addition of Postulates 4 and 5 the second theorem may be deduced.

Theorem 2. As the sound is continued, the raised visual threshold will tend to drop, in normal subjects.

Theorem 3 is based on the assumption that severe cortical lesions often interrupt the paths of communication from primary projection neurons to commonly recruitable neurons. Several possible ways in which this might occur follow:

1. Damage to, or extirpation of, cortical tissue might reduce the number of mutually recruitable neurons.

2. The location of the lesion may actually impinge upon and reduce the efficiency of neural pathways which offer the means of communication.

3. Edema associated with tissue damage might extend the area of altered tissue functioning beyond the site of damage.

4. It is frequently observed that in individuals with gross brain damage there is neural discharge at the periphery of the lesion. Such discharge could reduce the efficiency of communicating neural links if these paths were excited by the lesion discharge. That is, in summary neurons and neural functioning is impaired by brain damage. Postulate 7 includes these possibilities:

Postulate 7. There is a class of events D (which includes brain injury), which impairs process P.

Since neural communication will be less effective where tissue damage has occurred, the capturing of mutually recruitable neurons will be less, with accessory stimulation, than where undamaged tissue exists.

Theorem 3 follows from the inclusion of Postulate 7:

Theorem 3. The visual threshold of subjects possessing an element from the class of events D (brain injury) will be raised less than those of subjects free of events D, with the introduction of a moderately strong auditory stimulus.

A more extensive examination of the neurophysiological literature in support of the cited neural speculations will be presented in a later section of this paper.

Early Studies and Theoretical Contributions

The history of the problem starts with Urbantschitsch (1888, 1902). His is still one of the most extensive studies of this type for he proceeded to investigate all the sensory modes, and noted the effect of each upon the other. His results, however, were not very consistent and controls of extraneous physical conditions likely to influence the results were almost entirely lacking. The existence of intersensory effects, however, has been supported by a host of later investigators. Heymans in 1904 investigated the effect of electrical stimulation of the hand on auditory sen-

sitivity. Two subjects were used in a series of trials.

In one series the proportion of time a watch was heard at a given distance was used as a criterion of auditory sensitivity. Without electrical stimulation, the watch was heard almost all the time, but with increased intensity of shock the time during which the watch was heard was reduced, in some instances, to as little as 69 seconds in a 5-minute period. In another series, the distance threshold for auditory sensitivity under the same condition was tested. It decreased from about 2 meters without shock to about 1 meter with the most intense shock, which again indicated a decrease of auditory sensitivity.

A few years later, Jacobsen (1911) presented evidence to demonstrate that sound diminished the strength or intensity of weight sensation. This was investigated with a judgment of weights with and without a simultaneous sound. He also reversed the procedure and reported that sounds were judged to be louder without concomitant pressure sensations.

In contradiction to these demonstrations of sensory inhibition there are other very similar studies illustrating a facilitative effect of a secondary sensory stimulus upon a primary response. Ide (1919) tested the effect of temperature on weight judging. Both cold (45 degrees fahrenheit) and hot (147 degrees fahrenheit) weights were found to feel heavier than comparison weights at room temperature. Ide believed that the effect was simply a matter of increasing the total amount of sensation.

Hartman (1933) studied the effects of simultaneously presented stimuli upon an acuity threshold. His subjects judged the ease of discrimination of figures on a contrasting background (black on white and white on black). He reports evidence indicating that visual acuity can temporarily be increased by the simultaneous application of auditory, olfactory, and cutaneous stimuli, and that high and low tones, pleasant and unpleasant odors, mild tactile and pronouncedly painful stimuli, all enhance the ability to discriminate the test configurations.

These contradictory results indicate that heteromodal stimulation involves more than a determination of whether or not auxiliary stimulation is either inhibitory or facilitative. Gilbert (1941) has pointed out, in an article reviewing intersensory effects, that temporal and quantitative factors should be considered in examining heteromodal results. This article notes that Jacobsen recognized that a continued auxiliary stimulus could lose much of its inhibitory power and that it might, under certain conditions, augment rather than inhibit a primary response. Gilbert's conclusions, from an analysis of the studies reviewed, suggest a key to understanding the effects described and involve the following considerations. (*a*) A sufficiently intense stimulus will momentarily reduce sensitivity in another modality, and increase it after an optimum interval. (*b*) A less intense heteromodal stimulus will momentarily increase sensitivity.

Since this hypothesis allows an integration of results that must otherwise appear contradictory it is useful in the review of the investigations presented below.

In 1923, Newhall, and in 1934, Thorne, studied the effect of auditory stimulation on visual sensitivity with both stimuli presented simultaneously. In Newhall's study the subjects judged a given liminal light as superthreshold when clicks were added. Thorne's results were based on measurements of a visual threshold made under conditions of silence and a simultaneous buzzer. The effects of the buzzer were not constant; both facilitative and inhibitory effects were observed. On the whole, however, inhibitory effects were much more marked.

Thorne suggested that when the auxiliary stimulus was relatively strong it "becomes a figure in the perceptual figure-ground relationship and raises liminal sensitivity or exerts an inhibitory effect; when it continuously occupies the ground, it facilitates with resulting lowering of the threshold."

When we apply the Gilbert hypothesis to these two studies, we would expect Newhall's results to reflect a facilitative effect since a click must be a relatively mild auxiliary stimulus. Thorne's buzzer, on the

other hand, could be loud enough to produce the predicted inhibition. It is of interest to remember that Urbantschitsch reported that a loud tone or noise was necessary to achieve the inhibitory effect. In considering this, it is also of interest to note Jacobsen's stimulus conditions. He found that when pressure was inhibited by sound, weights of 10 to 30 grams were used; but when pressure inhibited auditory sensation, a weight of 300 grams was used as the stimulator of the auxiliary sensation of pressure.

The investigator who has produced the greatest amount of well-controlled quantitative data in this field is Kravkov (1930, 1936). He has investigated a wide variety of stimuli effecting a modification of the visual process. His results indicate that a concomitant auditory stimulation has a pronounced effect on both peripheral and foveal vision. He found that a tone of 2,100 cycles per second and 100 decibels greatly diminished light sensitivity for peripheral vision but increased sensitivity of foveal vision to white light (Kravkov, 1936). His studies also indicated that visual acuity could be altered with a concomitant auditory stimulus (Kravkov, 1930). His findings have been substantiated by Semenovskaia (1946), who found a similar effect under comparable conditions, and by Bogoslowski and Kravkov (1941) who found that the noise of an airplane motor raises the threshold of the rod apparatus in night vision, while in foveal and day vision the threshold is lowered.

The explanation offered for these effects is based on the concept of irradiation. Kravkov and his followers suggest that the excitation in the brain originating with the sound does not remain strictly localized, but is transmitted to neurons of the optic nerve, because of their anatomical proximity; thus subliminal excitation is created in the visual center.

In these studies strong accessory stimulation resulted in both facilitative and inhibitive effects. The latter is in agreement with Gilbert's analytical scheme, but the facilitation of foveal vision is contradictory, and will be examined in greater detail in a later section of this paper.

Child and Wendt (1938) experimented with the influence of a flash of light upon the audibility threshold. This threshold was determined by a tone of 165-millisecond duration. Their principal experimental variable was the time interval between the flash of light and the tone. When the light and the tone were simultaneous, or when the light preceded the tone by a ½ second or 1 second, there was a highly reliable increase in the frequency with which near-threshold tones were reported as heard. The maximum effect was found when the light preceded the tone by ½ second, or preceded it by 2 seconds; there was no consistent facilitating effect. It should be noted, however, that the auxiliary stimulus was a 2-degree circular patch of light of approximately 50 footcandles, one-tenth of a second in duration. Here again the intensity of the auxiliary stimulus is not great, and inhibition would not be expected from the application of the assumptions that we have considered in connection with the previous studies.

Child and Wendt's study sharpens the importance of very short-time relationships and permits the introduction of a central explanation of the effects that is consistent with the findings of investigators in neurophysiology.

Child and Wendt felt that their findings suggest a temporal summation of excitation in the central nervous system. Facilitation was interpreted as being due to the convergence of the two sets of impulses upon a final common path, and intervals which permit facilitation were interpreted to be a function of the relative latency and recruitment periods of the two converging excitations. One difficulty with this explanation is that the facilitating intervals are, in general, longer than would be expected from the work of such people as Hilgard (1933) and Wendt (1930) who found auxiliary stimulation could facilitate a reflexive response only when the time interval was less than 300 milliseconds. Also, in motor summation studies it has almost always been found that when the interval between stimuli is increased beyond the maximum interval that permits facilitation, an inhibitory effect is exhibited, and that the

magnitude of the inhibitory effect is often greater than that of the facilitating effect.

The history of the literature focuses attention on the need for exacting temporal investigation in heteromodal studies, and it points to the importance of intensity of stimulation in intersensory effects. However, no direct investigation of strong continuous accessory stimulation has been attempted. An examination of such influences is carried out in the present thesis.

The literature also attempts to relate specific behavior to neural functioning, to bridge a gap between what is usually thought of as two different levels of approach to the explanation of behavior. In recent years such attempts have achieved considerable sucess. Hebb's (1949) theory neural processes in behavior offers challenging possibilities. and the theoretical framework is complete enough to offer an explanation for a wide variety of psychological events. Köhler and Wallach (1944) have advanced a theory of cortical functioning based on the perception of simple figures. Klein and Krech (1952) have extended Köhler's theory to account for inter- and introindividual differences in behavior. As a result of their study investigating figural aftereffects, they have suggested that the "concrete behavior" of brain-injured individuals described by Goldstein and Schurer (1941), Werner (1940), and others may be viewed as instances of disturbed integration. They feel that such behavior may be attributed to an impaired communication process among different cortical areas.

Klein and Krech's study then further suggests that two stimuli which normally alter cortical functioning enough to affect a visual threshold, when presented simultaneously, would be less pronounced in brain-damaged subjects.

The study described in the following pages attempts to examine this possibility.

METHOD

Subjects

Two groups of subjects were used in this study; one organically "normal" and the other made up of patients with cerebral lesions of a severe nature.

The "normal" group consisted of attendants at the Boston Veterans Administration Hospital. These subjects were free of both known cerebral lesions and serious visual or auditory anomalies as determined by a routine physical examination required by the Veterans Administration of such employees.

The experimental group was drawn from the neurological wards at the Boston Veterans Administration Hospital. Subjects known to be free of serious auditory or visual deficiencies, but with known cortical damage, were selected.

The two groups were equated with the following criteria:

1. The age range included subjects from 18 to 45. Within this age group there is no known reason to expect any differences in general physiological functioning.
2. Only male subjects were used in order to avoid problems associated with sex differences.
3. Intelligence was roughly equated by limiting those selected to an IQ range of 85 to 120. The Stanford-Binet vocabulary test was used when a more extensive evaluation was unavailable. The patients usually had complete psychological analyses and these subjects were selected on the recommendation of the staff psychologist who had worked with the patient. Inasmuch as some brain-injured patients suffered from various degrees of aphasia, a vocabulary index of intelligence was not always felt to be valid, and thus more extensive test results were used.
4. Subjects who had had long experience at such occupations as radio code receiving or positions involving extreme visual acuity were excluded.

Apparatus

The apparatus consisted of an observation box, an adaptometer (NDR-C Model 2A), a signal generator with earphones, and recording equipment (Figure 2).

The observation box was fitted with a rubber-edged eye shield that served effectively as a head rest. At approximately 20½ inches from the eye of the observer a small red fixation light was placed in the center of the observer's visual field. At 2 degrees (visual angle) below this, the target light was placed. Its diameter was ½ inch. The surface of the target area was diffused by a translucent screen. The light source for the target area was a 6-volt filament lamp. The filament current was supplied from 110 ac lines, but a Sola constant voltage transformer was used to avoid line fluctuations before the current was dropped to 6 volts. The fixation light was supplied separately with the use of a variac.

A Hartline adaptometer with a 2-log-unit glass wedge was used to control the target light intensity. The wedge was automatically driven by a small high speed motor. A button which the subject held in his hand would instantaneously reverse the direction of the motor and thus the direction of the wedge. When the motor was turned

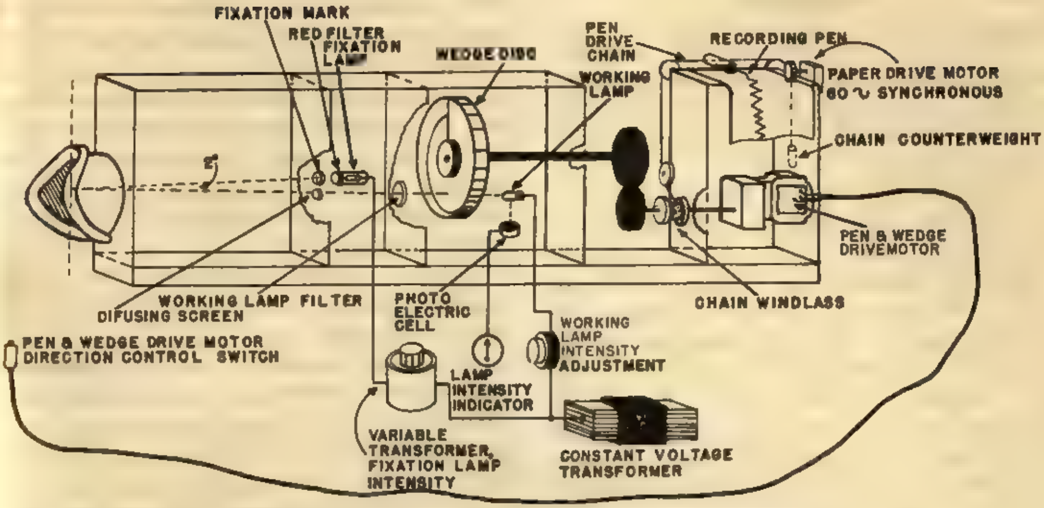

FIG. 2. Apparatus for the determination of visual thresholds.

on it ran continuously; only its direction could be altered.

A blue filter was placed between the wedge and the target area to neutralize the yellowish cast of the tungsten bulb.

In order to record the movement of the wedge a small drum was attached to its driving shaft, and a small chain, winding and unwinding on the drum, moved a recording pen back and forth over recording paper driven at 2.75 inches per minute.

The signal generator was an audio-oscillator producing a tone of 1,550 cycles per second. The earphones were type ANB-H-1 with sheepskin ear pads.

An adjustable chin rest was used to control head movement.

Procedures

All subjects were administered two or more drops of 1% euphthalamine, depending upon the amount needed to achieve an effective mydriasis. They were then dark adapted for ½ hour before the experimenting began. During this time the general procedure of the test runs was outlined to the subjects, and they listened for several moments to the sound used in the experiment in order to familiarize them with this stimulus.

The subjects were seated comfortably at a table with their heads supported by the head and chin rest on the adaptometer. In this position, two lights were visible, the red fixation light and the target patch. The subjects were told that the white light would gradually grow dimmer. They were instructed to push the response key the moment it reappeared.

A practice period of about 20 minutes was needed usually before a subject developed a reasonable proficiency in responding to the fluctuating light. All trials began with the target light at maximum intensity in order to provide a common starting intensity for all subjects.

Response speed was checked by suddenly covering the light source during a descending threshold determination. An immediate reversal of the recording pen, with this interruption, indicated an alert subject and a quick response. The completion of the practice series was followed by a short rest period. The glass wedge was then reset for maximum intensity of illumination and the testing started. The motor operating the adaptometer was turned on and the subsequent keying responses of the subject automatically recorded his thresholds. At the end of 2½ minutes a moderately loud tone of approximately 70 decibels (4 volts across earphones) and a frequency of 1,550 cycles per second was introduced by means of earphones placed on the head at the beginning of the experiment. The tone was maintained for 3½ minutes. At the end of this time the subjects were given a rest of 4 to 5 minutes. Two more 6-minute runs with an intervening rest period completed the series.

All threshold determinations were carried out in a completely darkened and reasonably silent room. An attempt to control the proper fixation by the subjects was made by reminding them to "watch the red light," a number of times during the series. It was always repeated a short time before the sound was turned on.

The equipment used was designed to reduce extraneous variables to a minimum. However, it is impossible to rule out such influences entirely. The comments of many of the subjects, in this study, indicated that the 6-minute period in which they constantly responded to the visual stimuli was too long and exacting to be a comfortable experience. There is no evidence to indicate that fatigue seriously influenced the results, but the patience of some subjects was apparently tried.

RESULTS

Analysis of the record. Figure 3 shows a section taken from the record of a normal subject. The plateau at the left traces the constant intensity of illumination at the onset of testing.

The first drop follows the decrement of illumination as the glass wedge rotates in front of the light source. The trough immediately following the drop indicates the time and intensity point at which the subject signaled the disappearance of the target. The subsequent rise traces the time and intensity during which the subject reported the stimulus light as absent. The following peak indicates the reported reappearance of the stimulus. The subsequent tracings repeat this cycle.

This record provides a series of ascending and descending thresholds comparable to the method of limits. In this case, of course, the light is a steady one and no discrete flashes are involved. The values used for computation were the distances of peaks or troughs, in millimeters, from the lower edge of the record. In this way low thresholds were represented by low values and high thresholds by high values.

The trough, in a record, was considered a measure of the last visible intensity for the subject, and the subsequent peak as a measure of the intensity at which the light reappeared for the subject. As in the method of limits, the mean of two such transition points was assumed to measure the momentary stimulus threshold. The peaks and troughs were combined in pairs, and the average for each pair was taken as a single determination.

The effect of sound. Figure 3 traces the visual thresholds of a normal subject for 6 minutes. The first 2½ minutes show the thresholds before sound is introduced, the next 3½ minutes show the thresholds while the subject is under the influence of the constant auditory stimulus. The point where sound is introduced is indicated by an "S" on the record.

In order to test the effects of the sound stimulus on the visual threshold, four 1-minute samples were drawn from each record. Such 1-minute periods provided equal temporal divisions, with only a few introductory and terminal thresholds disregarded. The first two samples were made up of threshold readings of the two 1-minute periods immediately prior to the introduction of sound. The third and fourth samples consisted of the threshold readings for the third and fourth minutes of sound. The thresholds for each time period were then averaged.

The averages for each of the four time periods in individual records were based on the same number of threshold readings. Usually the number of thresholds in each time period were the same. Occasionally a time period had two or three more threshold readings than another. In this case thresholds were numbered from the beginning of a time period and superfluous thresholds were rejected on the basis of the appearance of this number in a random number table.

Although the same number of thresholds were used in averaging time periods for a single individual, different records (inter-individual) varied in the number of thresh-

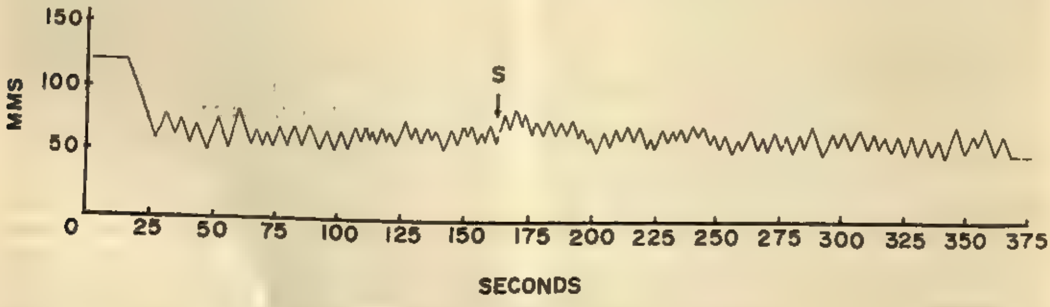

FIG. 3. Record of visual threshold determinations of a lesion-free subject.

TABLE 1

ANALYSIS-OF-VARIANCE THRESHOLD SCORES AND MEANS FOR NORMAL AND BRAIN-DAMAGED SUBJECTS FOR FOUR CONDITIONS

Source of Variance	*df*		*MS*	*F*
Normal *Ss*				
Total	91		—	—
Conditions	3		12.31	7.42**
Individuals	22		104.28	63.20***
Residual	66		1.65	—
Brain-damaged *Ss*				
Total	91		—	—
Conditions	3		3.29	1.60*
Individuals	22		212.35	103.58***
Residuals	66		2.05	—
Conditions	1	2	3	4
Means normal *Ss*	80.87	80.85	82.41	81.33
Means brain-damaged *Ss*	82.18	82.10	82.92	82.54

* $p > .05$.
** $p < .01$.
*** $p < .001$.

olds used in obtaining a time period average. This was due to individual variation in threshold determinations in the 1-minute periods, that is, some subjects indicated a large number of disappearances and reappearances of the light during this time than others.

Differences in thresholds for sound-on and sound-off periods were evaluated by separate analysis-of-variance procedures for the normal and brain-injured groups, with the time periods represented as four conditions (Table 1).

The variances associated with individuals and conditions were obtained from a two-way analysis of these tables. The F values in Table 1 indicate the significant differences. Individual differences are the most striking ($p < .001$) for both groups. The implication of such results, however, is hardly startling since individual differences in visual sensitivity are obvious to the casual observer. The difference between conditions is significant for normal subjects ($p < .01$), but insignificant for the brain-injured cases ($p > .05$). In both tables the means of Conditions 3 and 4 are higher, but the rise is less in the brain-injured group. These results indicate a significant rise in threshold for the normal group under the influence of sound and confidence limits placed on the means for conditions overlap, however, except for Condition 3. Thus, only the threshold rise for this period (1.56 millimeters) is significant.

The variance and means of the presound thresholds, included in the tables, are slightly higher in the brain-injured group. This suggested that the two groups might have been drawn from different populations of absolute thresholds. A t test of these thresholds, however, showed no significant difference ($p > .50$). The two groups were therefore assumed to be homogeneous with respect to absolute threshold, and differences in the two groups with respect to threshold changes were used for comparison. Since the analyses of variance only provided estimates of intragroup thresholds, t tests were used for between-group comparisons. Two postsound periods were selected for comparison with the minute period just prior to the introduction of sound. In the first case the first ½ minute of sound was used. The average difference of the two periods for normal subjects was compared with the average difference of these two periods for the brain-injured group. The resulting t value

of 1.77 lies between the .05 and .025 level of significance, and indicates a reliably greater rise for the normal group during this period. In the second case the first minute of sound was used. The results do not indicate a significant difference ($p < .13$). This suggests that a drop, after the initial rise, in the visual threshold is contributing to the average of this time period.

In order to determine whether or not the thresholds of normal subjects remained elevated as long as the sound continued, the differences between the first ½ minute and the third minute of sound for each individual was obtained. A t test applied to these values demonstrated a significant drop ($p < .01$).

Two problems remain in the analysis of these data. The first one is concerned with the rate of rise and fall in individual records. Figure 4 shows a curve roughly fitted to averaged thresholds for the normal group. The first 3 minutes of sound are represented. The threshold average for the 1 minute prior to the introduction of sound was used as the origin of the time axis. Nineteen points equally spaced along the time axis represent the points at which the records were sampled for threshold values. The question arises as to whether

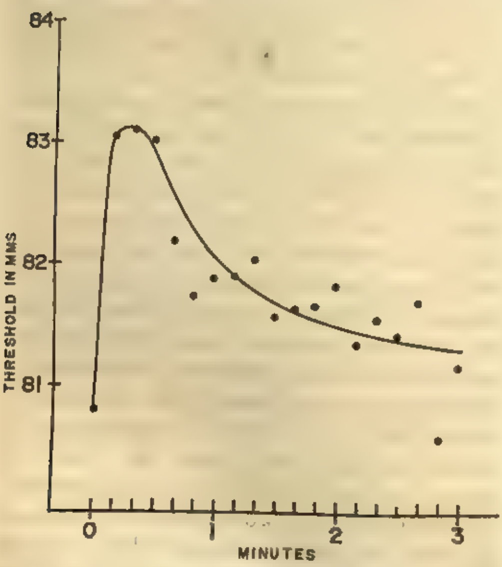

FIG. 4. Mean visual thresholds in relation to duration of auditory stimulation.

these values are the result of a relatively gradual change in visual threshold or are averages of momentary deflections. In order to answer this question a ranking scheme was resorted to. The first three threshold values after the introduction of sound were scored plus or minus to indicate an increase or decrease in threshold value from the immediately prior response. The number of positive values was tested against the number of positive values expected by chance alone. The results of a t test indicate a significantly higher number of positive values than would be expected by chance ($p < .02$), and therefore indicate a gradual rise in the visual threshold.

The second problem is concerned with the difference between the ascending and descending thresholds in the two groups. On any record, the distance between a peak and a trough is a measure of the time a subject takes before indicating the appearance or disappearance of the target light. An average of such times was obtained for each subject. A t test was then used to test the significance of differences between normal and brain-injured subjects. The resulting t value of 2.69 indicates a significant difference ($p < .05$). Either there is some perseverative influence in the perceptual experience of some patients, or their reaction times in terms of key pressing is significantly longer than for lesion-free subjects. In either event is seems unlikely that the results would be seriously influenced. In the first instance, it would seem probable that the delay of response was the same both for the disappearing and reappearing target. This would leave unaffected the mean value based on these two points. In the second instance, it might be argued that a perseverative effect operated in only one direction, that is, while the light was either on or off but not both. In this case the threshold means would either be raised or lowered. Such an argument, however, would have difficulty explaining the insignificant difference in the presound thresholds of the patient and normal groups.

Further comparisons. On the whole, nor-

mal records, when compared with those of brain-damaged cases, are characterized by a larger number of thresholds per minute and a greater rise with the introduction of sound. These factors, however, are not definite enough to be obvious by simple inspection. There is a considerable amount of overlap; many of the records of brain-injured subjects show a response frequency as high as normal controls, and there is an obvious overlap in threshold rise with the introduction of sound.

Possible vitiating effects. Introspective reports from many of the subjects indicated that the introduction of the sound was interpreted as an attempt to distract the observer and an effort was usually made to disregard the sound. It is impossible to evaluate the effectiveness of this attempt, however, and it is obvious that changes in threshold sensitivity may be brought about by involuntary fluctuations of attention. Perceptual discrimination of a near threshold light undoubtedly demands a precise adjustment of receptor and cerebral mechanisms, and a distraction or shift of interest to something outside the task would probably change this adjustment and alter the effectiveness of the performance.

However, a consideration of the temporal course of the raised threshold in the present study is of interest here. Once sound is introduced, the raised threshold shows a maximum inhibition of visual sensitivity some 15–18 seconds later. Since 4–5 seconds were used to increase the intensity of the sound to a maximum this leaves a "peak distraction" some 10–12 seconds after the introduction of the sound. Such an effect would require a concept of distraction with a rather long and relatively constant interindividual maximization. Usually distraction is purely a description designation which is chosen to signify a momentary disturbance that is somehow related to an interruption of assumed neurophysiological processes which are responsible for the organization of a particular type of behavior. The curve which describes the raised threshold in the present study has already been shown to represent a relatively gradual climb. Such a process is inconsistent with the conventional concept of distraction. The ascription of this term to this effect might well be quite proper, but the value of such an ascription without some specificity of process to account for the temporal span is obviously dubious.

Accessory stimulation may operate to produce its effect in still another way. The pupillary reflex may respond to autonomic innervation brought about by the relatively sudden noise. In the present investigation such a pupillary response was controlled by the use of the mydriatic euphthalamine. The effect of this drug is to dilate the pupil for several hours, preventing contraction to afferent stimuli. Consequently it seems reasonable to rule out an effect from this source.

Again, the performance of a given sense modality may frequently depend on associations aroused by the stimulation of another modality. In the complex perception of man this is an all-important factor. Associations developed over the years and the accumulated experience, which together become the basis of automatic unconscious deductions, are bound to play more or less decisive roles in the accessory action of one sense modality on another. It is believed that the simple physical properties used in the present study were innocuous enough to keep such determinants at a minimum.

The greater difference in threshold rise with the introduction of sound in the normal, as compared to the brain-damaged group, is significant at the .05 level. This is not a strong difference and suggests that the location of lesions in the experimental group should be related to heteromodal effects. The determination of the importance of such factors, however, is limited by the theoretical design of the investigation. In the present study no assumptions were made as to the exact area involved in intracortical communication, and thus no prediction as to the type of injury producing a minimum rise in the visual threshold was predicted. It was felt that the selection of any experimental group with lesions so restricted as to include only

areas of particular interest was virtually impossible. For this reason, an attempt was made to select cases with gross cerebral injury in the hope that critical locations would often be included.

A COMPARISON OF THEORIES

Traditionally, it has been assumed that the relation between stimulus and subjective intensity is a relatively constant function. Deviations from this relationship were explained in terms of errors of observation and the probability function of such errors. As these distributions have been examined more closely, such factors as practice, fatigue, changing metabolic processes, and a host of affective phenomena have been demonstrated as contributing to subjective effects.

A clarification of one of the psychophysical problems is necessary before the factors under investigation can be isolated for consideration. It is helpful if a differentiation is made between stimulus or physical intensity, excitation or physiological intensity, and subjective or psychological intensity. Recent developments in neurophysiology have enabled us to appreciate more thoroughly the neurological basis for the psychophysical relation. Studies of nerve action currents have thrown some light on the correlation between neural excitation with physical stimuli and the functional relationship of these two factors to a subjective response.

However, the measurement of the relation between stimulus and subjective intensities is complicated. Modern psychology recognizes this possibility and looks for explanations of behavioral processes in physical-physiological and physiological-psychological functions, realizing that each has its own determinants.

The studies reviewed here obviously deal only with physical stimuli and a behavioral response. In most cases the explanations either explicitly or implicitly recognize that the results obtained include a transformation of energy from physical stimuli to physiological processes and from these processes to psychological behavior.

The theories presented in the following pages attempt to construct a bridge between the physical stimulus and the behavioral response. Only the aspects relating to accessory stimulation are included. In this section, the attempt is made to consider, in a critical and constructive way, explanations of heteromodal effects that most closely adhere to and organize the behavioral data found in the literature and in the present study.

Holistic or organismic theories. General psychological theories have been developed in an effort to understand and to predict the day-to-day behavior of individuals in a variety of roles. Personality theorists, in particular, built with general concepts logical frameworks supported by success in prediction and therapeutic progress based on the theory. Experimentation at this level has led to the construction of concepts which are found necessary as building units. Such psychological elements are viewed as functioning in some type of patterned relationship. Usually, however, these elements and relationships have little applicability to the present problem, but the theories of Werner (1940) and Lewin (1936) are specific enough so that they may be used to predict some of the general results of this study. Although the explanations which are offered fail to cover all details, they were influential in determining the approach to the present study and they merit consideration.

Werner's theory of genetic development, and, more specifically, his treatment of psychological processes in brain-injured subjects, account for an effect in heteromodal stimulation. He has cited synesthetic data in support of this theory postulating a greater nonspecificity in function, in children. He has suggested a diffusion of sensory processes in the child which may gain greater specificity and organization with age. He postulates that damage to any part of the body affects the organism as a whole, reducing the effectiveness of the integrative processes. Werner would thus predict an altered visual threshold with the introduction of sound and altered effects with subjects suffering from brain damage. Lewin's reasoning follows the

same general lines (Lewin, 1936). Goldstein's experimentation (Goldstein & Schurer, 1941) has supported the possibility of a reduction in integration of psychological processes when injury to the brain is present. Again brain injury is seen to isolate psychological processes and thus predicts the reduced effect of sound in such cases.

A lack of exactness in holistic theories, in the area of heteromodal effects prevents a critical examination of the present data by these explanations. Intermodal influences are predicted along with smaller effects in brain-injured subjects, but no detailed account of temporal or intensity relationships are proposed.

Heteromodal theories. Facilitating effects with accessory stimulation were once looked upon as an example of the "dynamogenesis" of Fere (1887) and the "Bahnung" of Exner (1882). It was assumed that accessory stimulation resulted in a general physiological tonic effect. No attempt was made to specify the exact nature of this process.

Most experimenters now agree that the processes producing heteromodal effects are probably central ones. There is some variety in the specific cerebral functioning postulated, but it is usually specified or implied that a better understanding of the processes involved lies in the mechanisms of neural links.

The presence of intercentral connections makes such a consideration natural. The consensus of opinion involves the likelihood that an excitation originating in any of the receptors is not confined to a particular sensory modality but spreads to other areas of the nervous system.

Hartmann (1933), Kravkov (1936), and Child and Wendt (1938) assume a summation of afferent excitation in cases where the accessory stimulus enhances the tested modality. Their explanations are based on the neurological concept of "facilitation." It is assumed that if two afferent volleys are small, each may result in the discharge of impulses by a certain number of neurons while exciting others to a degree short of that required to produce discharge of impulses; in other words subliminally. Under such circumstances, the field occupied by neurons excited to discharge is called the "discharge zone," while the field of neurons receiving subliminal excitation is called the "subliminal fringe." When two nerves are stimulated at the same time, the resulting discharge may be greater than when fired individually. It is presumed that, at some place in the brain, a "subliminal fringe" of the accessory excitation overlaps with a "subliminal fringe" of the tested modality. Thus, when the two senses are simultaneously stimulated, there is an enhancement of effects. Hartmann (1933) has placed the area of overlapping "subliminal fringes" in the cortex while Kravkov (1936) speculated that since the auditory and visual pathways are proximate in the corporaquadrigemina, this area is a likely point of convergence. Child and Wendt (1938) do not attempt to localize this process but only point out that their results support the likelihood of heteromodal "facilitation." They found a maximal effect of an accessory light stimulus on the auditory threshold when the two stimuli were presented simultaneously or within ½ second of one another. These time intervals are shown to approximate those of the motor summation studies of Bowdich and Warren (1890), who studied the time relationships of the summating influences of a variety of stimuli upon the patella reflex.

Kravkov (1930) explains an experiment on visual acuity in this way. Because of "irradiation" light objects against a black background take on larger size. Kravkov also suggests, "an additional excitation of the brain can be produced by different indirect stimuli such as illumination of the other eye, a tone, or an odor. In the latter cases the excitation of the neurons directly involved is transmitted in the neurons of the optic nerve, thanks to their anatomical proximity; and thus a subliminal excitation is created in the visual center." In this way an increase of irradiational effect would assist discrimination by making larger the white interspaces between black objects crowded on a white background.

For the same reason the irradiational effect would make more difficult the discrimination of small white objects crowded together against a black background. The black interspaces would be reduced, to the perceiving eye, discrimination would thus be hampered, and visual acuity would suffer accordingly. However, the theory cannot adequately explain Kravokv's results which deal with the effect of sound on peripheral vision. Here visual sensitivity was shown to be reduced and Kravkov can only hypothesize intercentral connections that "dominate antagonistically over facilitating influences [Kravkov, 1936]."

Although most of the recent investigators turn to the central nervous system for an explanation, Thorne (1934) suggests that the results be considered in phenomenological terms. He made measurements of the visual threshold under conditions of silence and a simultaneous buzzer. He found both facilitative and inhibitory effects, with inhibition much more marked. His explanation is based largely on his own introspection as he served as a subject. He suggests that when the auxiliary stimulus is relatively strong, it "becomes a figure in the perceptual figure ground relationship and raises liminal sensitivity or exerts an inhibitory effect; when it continuously occupies the ground it facilitates with resulting lowering of the threshold."

Field theory. Gilbert (1941) has turned to Köhler's electrical field theory to explain altered receptor processes. Köhler and Wallach (1944) conceive of certain cerebral areas as responding to electrical stimulation in much the same way as physical volume conductors do. It is assumed that a current passing through a given area increases the resistance of this pathway, thus inhibiting the immediately following excitation. Subsequent currents tend to be shunted into adjacent tissue. This alteration in the locus of the electrical potential is assumed to be responsible for a distortion of perceptual discrimination. The resistance is termed "polarization," and it is used, in the physical sense to describe the production of an electromotive force acting in the opposite direction to an original current. Gilbert tentatively suggests that lines of force originating with excitation from one modality, are hampered in the shunting process by antagonistic lines of force which are created by the introduction of the accessory stimulus. This accounts for the decrement of subjective intensity in the primary modality. Gilbert extends this explanation to account for facilitative effects. He suggests that an accessory stimulus of moderate intensity, introduced and removed at the proper times, might facilitate the irradiating lines of force from the first stimulated area. Although no details of this process are given by Gilbert, he implies the use of superexcitability which is believed to be a momentary effect of the collapsing accessory field. As the accessory field collapses, the momentary flow of the counter electromotive force developed by polarization, would aid the invasion of irradiating lines of force from the primary modality.

Such an explanation, although it accounts for both the inhibitory and the facilitative effects, lacks the specificity which is needed to explain the decrementing auditory effects in this study and makes no attempt to deal with exacting temporal relationships.

Studies involving brain-injured subjects. Controlled studies involving heteromodal effects in brain-injured subjects are conspicuously absent in the literature. Altered visual effects in subjects suffering from severe cerebral lesions, however, have been demonstrated. Werner and Thuma (1942) have compared birth (brain)-injured children with children who made low scores on a standardized test for intelligence and found critical flicker fusion to be reliably lower on the average in the former. They noted that the difference was most marked at low intensity levels of the intermittent light and virtually disappeared as the intensity of the flashes was maximized. Halstead (1947) found a similar effect of low level intensity in brain-injured patients. He interpreted his findings as providing important clues as to the nature of the processes reflected by critical-fusion frequency: "Here for the first time we have direct

evidence that they (c.f.f.) are central (cerebral) processes rather than peripheral (retinal) as they have traditionally been regarded." In addition, Halstead and his associates (1947) have found that the dominant brainwave rhythm in the monkey electroencephalogram can be driven up to the point of critical-fusion frequency only, and the visual pathway below the cortex can be activated by photic stimuli to the retinas at suprafusional values. This fact is in line with the evidence presented so far which links altered visual thresholds with cerebral processes.

Klein and Krech (1952) have not only adduced data supporting a central explanation of altered sensory processes in brain-damaged subjects, but they have offered a theory that allows an interpretation of heteromodal effects.

Klein and Krech have demonstrated that as an individual is continuously subjected to kinesthetic stimuli, a decrease in the subjective intensity of this stimulus is experienced. They have interpreted these results as evidence of Köhler's "polarization." Since any afferent stimuli would presumably follow a similar decline, any process demonstrating intracortical communication would soon show a decrement in effects. Thus, general results of the normal group in the present investigation are implicitly predicted. In addition, however, Klein and Krech offer evidence that suggests that polarization is still increasing at the end of 2 minutes of stimulation. The curve, tracing the effects of sound, in the current study is still dropping and approaching the presound threshold at the end of this time. Again, such an explanation predicts reduced transmission of excitation through the brain field in brain-damaged subjects. Polarization, which is assumed to inhibit communication among localized cortical regions, is more pronounced, and occurs after less prolonged stimulation in brain lesion cases, as measured by kinesthetic figural aftereffects. In this way, intracortical communication could be influenced earlier and more intensely in the group.

Klein and Krech thus account for the decline, after the initial rise with sound, of the visual threshold described in this paper. They also account for the smaller rise in brain-injured subjects.

They do not, however, account for the initial influence of one sensory modality upon another; and whether or not their theory would adequately cover the effects of sound on the visual threshold, at the end of three minutes of stimulation must await further experimentation.

BRAIN MODEL AND NEUROLOGICAL PROCESSES

The background for the postulates in the introductory section of this paper was described with little authentication for the neural processes presented as characteristic of a brain model. This section will offer documentation and support for the plausibility of such thinking.

The physical aspects of signals entering the central nervous system from a variety of end organs have been directly examined. These include studies of visual sense cells by Hartline and Graham (1932), muscle-stretch receptors by Matthews (1933), sense organs for taste by Pumphery (1935) and Pfaffmann (1941), pressure receptors in the carotid sinus by Bronk and Stella (1932) and pain fibers in the skin by Adrian (1928) which points out that two fundamental facts are evident:

1. Though the primary stimuli to the sense organs are physically different, the signals communicated to the nervous system are alike in that they consist of a train of action potentials.

2. An increase in the intensity of the stimulus is reflected in an increase in the frequency of transmission of impulses into the central nervous system without significant change in the magnitude of the individual potentials.

With such evidence the first postulate describing a characteristic of the model in Figure 4, is in accord with current neurological evidence. It states that the frequency of neural discharge in a modal network is positively related to the strength of the sensory stimulus. Only the generality of the statement requires more support.

Since the study described here used a rather definite auditory intensity, it is important to examine, critically, the auditory action potential at that intensity (approximately 70 decibels). Stevens and Davis (1938) have shown, that, as the intensity of an auditory stimulus is increased, both the cochlear microphonics and the action potentials grow larger. Near threshold the increases can be measured satisfactorily, but at 30 or 40 decibels above threshold, measurement becomes difficult because the mechanical response of the ear to the stimulus is not critically damped, and the action potentials are superimposed on the later waves of the microphonics. Precise measurements are therefore impossible at high sound intensities, but it can be seen that the action potential continues to increase as well as the cochlear microphonics, although they do not necessarily follow the same law of increase. These findings add support to the first postulate and its applicability to the present study.

It is also presumed (Postulate 4) that with a constant stimulus the frequency of neural discharge in its modal network does not remain constant but reduces to approach a stable lower level as the stimulus is maintained. Again evidence for the complete generality of the assumption is lacking. With auditory stimulation, however, the relationship cannot be doubted. It has been demonstrated by measurement of the action potential of the auditory nerve that, immediately after the introduction of a continuous tone, the action potential does not remain constant in size but shrinks, first rapidly, then more slowly to a lower amplitude. The same drop was found by Galambos and Davis (1943) who found both a rate adaptation and an amplitude adaptation. They point out that the functional refractory period of the auditory nerve is not always constant. It increases significantly as stimulation is continued, and the threshold of stimulation for each fiber tends to rise also. Since the action potential is known to depend upon the number of individual fibers involved, it is assumed that such equilibration is the result of a dropping out of a number of fibers. Stevens and Davis (1938) suggest that reduction processes, extending over several minutes, represent a readjustment of the chemical dynamics in the nerve fiber and the attainment of a new equilibrium between anabolism and catabolism. With an intensity and frequency paralleling that used in the present study, they demonstrated that the shrinkage of the action potential continues for 5 to 7 minutes after the onset of the stimulating tone.

It is assumed in Postulate 5 that interruption of a continuous stimulus would permit neural recovery and tend to eliminate the effects of tetanic stimulation.

The nonspecificity of neurons is postulated with no real neurological support. Certain effects of sensory impulses, however, lend themselves to speculation in this direction. Afferent stimulation has been demonstrated to lead to as many as three distinct responses in the cerebral cortex. The responses may be distinguished on the basis of latency, threshold, and localization in the cortex. In 1937, Marshall, Woolsey, and Bard recorded a latency of 8 to 10 milliseconds to afferent stimulation in the monkey, and the effect was sharply localized in the sensory cortex. In 1939, Forbes and Morrison (1939) obtained two cortical responses to stimulation of the cat's sciatic with single shocks. An initial "primary" response whose latency was 8 to 10 milliseconds, was followed by a "secondary" response 30 to 80 milliseconds after the stimulus. The secondary response was obtained equally well in all regions of the cortex on both the contralateral and ipsilateral sides. Heinbecker and Bartley (1940) describe cortical responses whose latencies are similar after stimulation of the saphenous nerve in unanesthetized cats. In addition they describe a third response of still longer latency (400 milliseconds) which occurs only after stimulation strong enough to activate the "C" fibers of the nerve.

Secondary discharges similar to those under discussion have been elicited by various types of afferent stimulation. The

tactile sensory study of Marshall, Woolsey, and Bard has already been mentioned. Bishop (1936) has reported similar "nonspecific" responses with optic excitation, and even labyrinthine stimulation has resulted in similar effects as reported by Gerebetzoff (1940). Bremer (Bremer & Bonnet, 1950) reported in 1950 the occurrence of a long latency response in the visual cortex to auditory stimulation that was introduced in the form of clicks. His observations have been more recently confirmed by Thompson and Sindberg (1960). Again Buser (Buser, Bruner, & Sindberg, 1963) has suggested a relatively specific location for such interaction. The relationship between this striking type of electrical activity in the cortex and the function of the cortex is as yet unknown. Although no specific evidence demonstrates the involvement of the same neural chains, and parallel neural systems may account for this "nonspecific" response, the assumption that the same system of neurons is aroused by two different sensory modalities is now a common hypothesis.

Postulate 3 states the conditions under which the hypothetical mutually recruitable neuron shifts its influence from one modality to the other. Such a switching mechanism has been suggested by Gasser (1937) in order to explain reciprocal inhibition, that is, inhibition of a motor neuron by kinesthetic impulses from another motor neuron anatomically near it. In such a response it is assumed that an internuncial neuron takes on the properties of a final common path. This neuron becomes a switch which determines which of two competing reflexes will have the right of way. Whichever way the internuncial neuron goes, so goes the reflex. It is supposed that the final common path may be taken over first by one stimulus and then by the other, so that the two stimuli take turns at calling forth their respective reflexes. One of the factors deciding which of two competing reflexes occur is the relative intensity of their respective stimuli. Other things being equal, the stronger stimulus will determine which of the reflexes will occur. Postulate 3 states this relationship of competing stimuli, and relates the recruitment of an internuncial neuron to intensity of stimulation. In this case, of course, the competing stimuli are assumed to stem from two different sensory systems rather than from two different responses to the same stimuli.

Postulate 6 relates neural phenomena to sensory experience. It can only be claimed at present that such physiological processes are correlated with stimuli potentially experiential. It is not known where "vision" or "audition" are consummated, but neural discharge must be a preliminary to them.

In evaluating the validity of the neural processes involved in this brain model, the threshold relationship to time is of critical interest. As already pointed out under the discussion of "attention", the rise, and subsequent fall, of this threshold is a relatively gradual process. The maximum rise for the normal group is not immediate but occurs some 10–20 seconds after the introduction of sound. It then declines toward the presound value, approaching it at the end of 3 minutes. It is of course assumed that with the onset of sound, neurons sensitive to excitation from either the visual or auditory pathway, are responding to the auditory stimulus and reducing the subjective visual experience. As already pointed out, the auditory pathway fatigues under tetanic stimulation, and a decline in the effectiveness of the auditory stimulus was for this reason, predicted. Galambos and Davis (1943) have demonstrated that auditory nerves respond to a continued stimulus of constant intensity by a burst of impulses which gradually decline in rate. This "rapid equilibration" occurs within a second or two. Derbyshire and Davis (1935), however, have demonstrated that the amplitude of the action potentials at the end of 2 seconds of stimulation is only a relatively constant value. If photographs are taken at appropriate intervals during 10 minutes following the first two seconds of stimulation, the amplitude shows the further reduction illustrated in Figure 5.

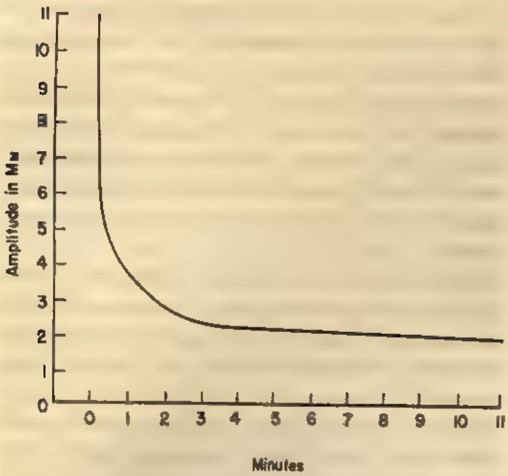

FIG. 5. Action potential of the auditory nerve in relation to duration of stimulation. (After Derbyshire & Davis, 1935.)

The minimum size of the response is attained in 7 to 10 minutes after the onset of the stimulus. Although Figure 5 traces the fall of potential at a frequency of 875 cycles per second, Derbyshire and Davis report that the degree and rate of equilibration is the same at 1,600 cycles per second. This is consistent with the theory that individual nerve fibers respond alternately when pacing ceases.

The decrement of the auditory potential over this time period, closely parallels the rate of drop demonstrated by the visual threshold in the present investigation. Figure 4 shows a smoothed curve of the averaged thresholds for the normal group. Once the threshold has maximized it can be seen to follow a rate of decline similar to that of the action potential represented in Figure 5.

The correspondence in these curves may be fortuitous, but it does offer evidence demonstrating that equilibration processes continue for periods consonant with the explanations offered by the brain model of the present study.

The rather slow rise in the visual threshold is not very well explained by the brain model. It can only be assumed that sometime, far longer than would be expected from direct synaptic mechanisms, is needed to innervate the processes resulting in the raised threshold.

Facilitation. The brain model offered in the present study allows for both inhibitive and facilitative processes with accessory stimulation. It proposes that stimuli of low intensity result in "facilitation," and that stimuli of high intensity result in inhibitory effects.

Although simple monosynaptic pathways exist in the central nervous system, most of the central reactions are mediated through neuron linkages of great complexity. Neurological theory (Fulton, 1949) assumes that when a volley of impulses impinges upon a pool of quiescent neurons, some of those neurons are excited to discharge impulses; others receive only subliminal grades of excitation; still others remain quiescent. The subliminal fringe forms the liaison between the discharge zone and the quiescent members of the neuron pool. That is to say, if the volley is increased, neurons are recruited from the subliminal fringe into the discharge zone, while others from the quiescent pool enter the subliminal fringe; if the volley is decreased, neurons from the discharge zone are shed into the subliminal fringe and other neurons from the subliminal fringe retire into the quiescent pool.

This process traces the phases through which a mutually recruitable neuron is presumed to pass. With a relatively mild auditory stimulus, such a neuron is included in a "subliminal fringe," thus facilitating the visual process; but as the intensity increases this neuron is captured by the auditory modality; finally it is again lost to the auditory system as auditory equilibration sets in and the volley of afferent impulses decreases.

In general the results of this study suggest basic processes of cerebral integration. Certain limited predictions based on early results of this study have been supported by specific experimentation. For example, changes in critical flicker fusion may be predicted. One of the variables in critical flicker fusion is the intensity of the flashing light. A reduction of light intensity lowers the fusion rate. Since it has been demonstrated that sound will reduce the subjective intensity of a light stimulus, it could be expected to lower the fusion

threshold if introduced while fusion determinations were being run. Gorrell (1953) has adduced data supporting this prediction. He has also found that sound is less effective in lowering the fusion rate of brain injured subjects. A comparison of adults and children has demonstrated a significantly different alteration in fusion rate under the sound stimulus.

The studies of Grüsser and Grüser-Cornehls (1960) lend general support to the assumed underlying mechanisms proposed in this thesis. They have demonstrated that the critical flicker frequency for single neurons (CFF, the maximum frequency at which neurons can follow flickering light) can be altered by concomitant vestibular stimulation. It seems plausible that such processes underlie the subjective flicker fusion experiences just described.

Theoretically any marked stimulation would be expected to lower sensory thresholds. On the basis of this reasoning severe anxiety and moderately loud sound may be viewed as producing functionally equivalent effects. That is, anxiety may, through the innervation of the sympathetic nervous system, result in marked thalamocortical firing which could interrupt neural patterns associated with sensory processes.

A demonstration of the effects of such physiologically defined anxiety would support an assumption of generality in bodily defense system, and it would aid in detecting the processes involved in sensory organization. In this sense, the present study brings within the scope of personality theory phenomena which were not previously regarded as particularly pertinent to it.

SUMMARY

The study of the processes underlying the transmission and coordination of two different kinds of sensory excitation in the cortex lies within the province of both the psychologist and the neurophysiologist. Both disciplines are examining factors contributing to the consummation of such integrated experiences. In the present study a neurological model accounting for specified heteromodal effects was proposed. The method used involved the determination of visual thresholds in normal and brain-injured subjects while they were exposed to an aural stimulus of moderately loud intensity. The results demonstrated group differences in the influence sound has on visual thresholds, and provided information on the diminishing effectiveness of a constant auxiliary stimulus when it is maintained for a period of several minutes. The results support the occurrence of intersensory effects and the probability of a centrally located mediating process. The introduction of sound initially raised the visual threshold of normal subjects and lost its effectiveness when maintained for a period of 3 minutes. The visual threshold of subjects in the brain-injured group were not raised as much as were the visual thresholds of the lesion-free subjects, with the introduction of a 1,550-cycle-per-second tone at approximately 70 decibels above threshold.

Possible vitiating effects such as distraction, pupillary reflex, and the meaningfulness of stimuli were discussed.

REFERENCES

ADRIAN, E. D. *The basis of sensation.* London: Christophers, 1928.

ADRIAN, E. D. *The mechanisms of nervous action.* Philadelphia: Johnson Foundation Lectures, University of Pennsylvania, 1932.

AMASSIAN, V. E. Studies on organization of a somesthetic association area, including single unit analysis. *Journal of Neurophysiology,* 1954, **17,** 39–58.

AMASSIAN, V. E., & DEVITO, R. V. Unit activity in reticular formation and nearby structures. *Journal of Neurophysiology,* 1954, **17,** 575–603.

BAUMGARTEN, R., VON MOLLICA, A., & MORUZZI, G. Influence of the motor cortex on the spike discharges of the bulbo-reticular neurons. *Electroencephalography and Clinical Neurophysiology, Amsterdam,* 1953, Suppl. 3, 68.

BISHOP, G. H., & O'LEARY, J. Components of the electrical response of the optic cortex of the rabbit. *American Journal of Physiology,* 1936, **117,** 292–308.

BOGOSLOWSKI, A. I., & KRAVKOW, C. B. The action of the noise of the airplane motor on vision. *problemy̆ Fiziologicheskoi Optiki,* 1941, **1,** 69–75

BOWDITCH, H. P., & WARREN, J. W. The knee-jerk and its physiological modifications. *Journa of Physiology,* 1890, **11,** 25–64.

BREMER, F., & BONNET, V. Interprétation des re

actions rhythmiques prolongées des aires sensorielles de l'écorce cérébrale. *Electroencephalography and Clinical Neurophysiology, Amsterdam,* 1950, **2**, 389–400.

Bronk, D. W., & Stella, G. Afferent impulses in the carotid sinus nerve. I. The relation of the discharge from single end organs to arterial blood pressure. *Journal of Cellular and Comparative Physiology,* 1932, **1**, 113.

Buser, P., Bruner, J., & Sindberg, R. Influences of the visual cortex upon posteromedial thalamus in the cat. *Journal of Neurophysiology,* 1963, **26**, 677–691.

Chang, H.-T. Cortical response to stimulation of lateral geniculate body and the potentiation thereof by continuous illumination of retina. *Journal of Neurophysiology,* 1952, **15**, 5–26.

Child, I., & Wendt, G. R. The temporal course of the influence of visual stimulation upon the auditory threshold. *Journal of Experimental Psychology,* 1938, **23**, 109–127.

Derbyshire, A. J., & Davis, H. The action potentials of the auditory nerve. *American Journal of Physiology,* 1935, **113**, 476–504.

Exner, S. Zur Kenntnis von der Wechselwirkung der Errengungen im zentral nerven System. *Pflügers Archiv für die gesamte Physiologie des Menschen und der Tiere, Berlin,* 1882, **28**, 487–507.

Fere, C. *Sensation et movement.* Paris: 1887.

Fessard, A. The role of neuronal networks in sensory communication within the brain. In W. Rosenblith (Ed.), *Sensory communication.* New York: Wiley, 1961. Pp. 585–606.

Forbes, A., & Morrison, B. R. Cortical response to sensory stimulation under deep barbiturate narcosis. *Journal of Neurophysiology,* 1939, **2**, 112–128.

Fulton, J. F. *Textbook of physiology.* Philadelphia: Saunders, 1949.

Galambos, R., & Davis, H. The response of single auditory-nerve fibers to acoustic stimulation. *Journal of Neurophysiology,* 1943, **6**, 39–58.

Gasser, H. S. Reciprocal innervation. In *Jubille,* volume in honor of Professor J. Demoor. Liege: G. Thone, 1937. Pp. 212–218.

Gerebetzoff, M. A. Recherches sur la projection corticale du labyrinthe; des effects labyrinthique sur l'activite electrique de l'ecorce cerebrale. *Archives of International Physiology,* 1940, 50, 59–99.

Gilbert, G. M. Intersensory facilitation and inhibition. *Journal of General Psychology,* 1941, **24**, 381–407.

Goldstein, K., & Schurer, M. Abstract and concrete behavior. An experimental study with special tests. *Psychological Monographs,* 1941, **53**(2, Whole No. 239).

Gorrell, R. The effect of extraneous auditory stimulation on critical flicker frequency. Unpublished doctoral dissertation, Clark University, 1953.

Griisser, O. J., & Griisser-Cornehls, U. Mikro-Elektrodenunterschungen zur Konvergenz vestibularer und retinaler afferenzen an einzelnen Neuronen des optischen cortex der Katzer. *Pflügers Archiv für die gesamte Physiologie des Menschen und der Tiere,* 1960, **270**, 227–238.

Halstead, W. C. *Brain and intelligence.* Chicago: University of Chicago Press, 1947.

Hartman, G. W. Changes in visual acuity through stimulation of other sense organs. *Journal of Experimental Psychology,* 1933, **16**, 383–392.

Hartline, H. K. & Graham, C. H. Nerve impulses from single receptors in the eye. *Journal of Cellular and Comparative Physiology,* 1932, **1**, 277.

Hebb, D. O. *The organization of behavior.* New York: Wiley, 1949.

Heinbecker, P. & Bartley, S. H. Action of ether and nembutal on the nervous system. *Journal of Neurophysiology,* 1940, **3**, 219–235.

Heymans, G. Untersuchungen uber psychiscke. Hemmung: V. Die Verdrangung von Schallemphfindungen durch elektrische Hautempfindhgen. *Zeitschrift fur Psychologie mit Zeitschrift fur Angewandte Psychologie und Charakterkunde,* 1904, **34**, 15–28.

Hilgard, E. R. Reinforcement and inhibition of eyelid reflexes. *Journal of General Psychology,* 1933, **8**, 85–113.

Ide, A. L. The influence of temperature on the formation of judgments in lifted weight experiments. Unpublished doctoral dissertation, University of Pennsylvania, 1919.

Jacobson, E. Experiments on the inhibition of sensations. *Psychological Review,* 1911, **18**: 24–53.

Jung, R. Neuronal integration in the visual cortex and its significance for visual information. In W. Rosenblith (Ed.), *Sensory communication.* New York: Wiley, 1961. Pp. 627–674.

Jung, R., Kornhuber, H. H., & DaFonseca, J. S. Multisensory convergence on cortical neurons. In G. Moruzzi, A. Fessard, & H. Jasper (Eds.), *Brain Mechanisms.* Vol. 1. Amsterdam: Elsevier, 1963 Pp. 207–240.

Klein, G. S., & Krech, D. Cortical conductivity in the brain-injured. *Journal of Personality,* 1952, **21**, 118–148.

Köhler, W., & Wallach, H. Figural after-effects. *Proceedings of the American Philosophical Society,* 1944, **88**(4).

Kravkov, S. W. Ueber die Abhangigkeit der Sehscharfe vom Schallereiz. *Archiv für Ophthalmologie,* 1930, **124**, 334–338.

Kravkov, S. W. The influence of sound upon the light and color sensitivity of the eye. *Acta Ophthalmologica, Copenhagen,* 1936, **14**, 348.

Lashley, K. S., Chow, K. L., & Semmes, J. An examination of the electrical field theory of cerebral integration. *Psychological Review,* 1951, **58**, 123–126.

Lewin, K. *A dynamic theory of personality.* New York: McGraw-Hill, 1936.

Licklider, J. C. R. On psychophysiological models. In W. Rosenblith (Ed.), *Sensory communication.* New York: Wiley 1961 Pp. 49–72.

Marshall, W. H., Woolsey, C. N., & Bard, P.

Cortical representation of tactile sensibility as indicated by cortical potentials. *Science*, 1937, **85**, 388–390.

Matthews, B. H. S. Nerve endings in mammalian muscle. *Journal of Physiology*, 1933, **78**, 1.

Newhall, S. M. Effects of attention on the intensity of cutaneous pressure and on visual brightness. *Archives of Psychology*, 1923, **61**, 75.

Pfaffmann, C. Gustatory afferent impulses. *Journal of Cellular and Comparative Physiology*, 1941, **17**, 243.

Pumphery, R. J. Nerve impulses from receptors in the mouth of the frog. *Journal of Cellular and Comparative Physiology*, 1935, **6**, 445.

Scheibel, M. E., Scheibel, A. B., Mollica, A., & Moruzzi, G. Convergence and interaction of afferent impulses on single units of reticular formation. *Journal of Neurophysiology*, 1955, **18**, 309–331.

Segundo, J. P. & Machne, X. Unitary responses to afferent volleys in lenticular nucleous and claustrum. *Journal of Neurophysiology*, 1956, **19**, 325–339.

Semenovskaia, E. N. Light sensitivity of central and peripheral vision as affected by acoustic stimuli. *Problemy Fiziologicheski Optiki, Moskva*, 1946, **3**, 94–96.

Stevens, S. S., & Davis, H. *Hearing*. New York: Wiley, 1938.

Thompson, R. F., & Sindberg, R. M. Auditory response fields in association and motor cortex of cat. *Journal of Neurophysiology*, 1960, **23**, 87–105.

Thorne, F. C. The psychological measurement of the temporal course of visual sensitivity. *Archives of Psychology*, 1934, 127.

Urbantschitsch, V. Ueber den Einfluss der Sinniserregung suf die ubrigen Sinnesempfindugen. *Pfluegers Archiv für die gesamte Physiologie des Menschen und der Tiere, Berlin*, 1888, **42**, 154–182.

Wendt, G. R. An analytical study of the conditioned knee-jerk. *Archives of Psychology*, 1930, **123**, 97.

Werner, H. *The comparative psychology of mental development*. New York: Harpers, 1940.

Werner, H., & Thuma, B. D. Critical flicker-frequency in children with brain injury. *American Journal of Psychology*, 1942, **55**, 394–399.

(Received May 30, 1966)